

Identification of Potential Drug Targets in Kinetic Networks Described by Ordinary Differential Equations

Inauguraldissertation

zur Erlangung des Grades eines
Doktors der Naturwissenschaften

am Fachbereich Mathematik und Informatik
der Freien Universität Berlin

vorgelegt von
Marvin Schulz

Berlin 2013

Erstgutachter: Prof. Dr. Alexander Bockmayr
Zweitgutachterin: Prof. Dr. Dr. h.c. Edda Klipp

Tag der Disputation: 14.12.2012

Preface

The first part of this thesis which is described in chapter 3 has been published in *Molecular Systems Biology* under the title “Retrieval, alignment, and clustering of computational models based on semantic annotations” [Schulz et al., 2011]. My contributions to this work have been the development and adaptation of similarity measures, their evaluation, and their implementation in the web tool semanticSBML, which has been published in *Genome Informatics* and *Bioinformatics* [Schulz et al., 2006, Krause et al., 2010].

The second part of the thesis, extending the similarity measures by incorporating model network structures, has been published in *BMC Bioinformatics* under the title “Propagating semantic information in biochemical network models” [Schulz et al., 2012]. For this work I have developed the “annotation propagation” method. Furthermore, I have implemented and evaluated the annotation and the similarity propagation method.

The third part of the thesis has partially been published in *BMC Bioinformatics* under the title “Tide: a software for the systematic scanning of drug targets in kinetic network models” [Schulz et al., 2009]. My contributions have been the development of the software for the identification of potent drug targets in kinetic models and its application to the glycolysis in *Trypanosoma brucei*. Further conceptual changes in the drug target identification framework as well as its application to the arachidonic acid pathway in humans have not been published so far.

Acknowledgements

I would like to express my sincere gratitude to my supervisors Prof. Dr. Dr. Edda Klipp and Prof. Dr. Alexander Bockmayr who have laid out the direction of this research, been a constant inspiration, and provided a helpful and supporting scientific and personal environment.

A significant portion of this work has been influenced by Dr. Wolfram Liebermeister, who has contributed many important ideas and has shaped the way in which I approach science in a sustainable manner. Thanks, Wolf.

Apart from people in Berlin different collaboration partners have in one way or the other contributed to my development and the work presented in this thesis. I would like to acknowledge Dr. Barbara Bakker, Dr. Klaas Krab, and Prof. Dr. Hans Westerhoff, who have provided a fruitful working environment in Amsterdam, and Dr. Dagmar Waltemath, Ron Henkel, and Dr. Nicolas Le Novère, who contributed significantly to our model similarity measure and our software semanticSBML.

The development of the ability to work productively in a team can be a long and rocky road. Therefore, I would like to thank all the people who have developed software together with me: Falko Krause, Timo Lubitz, Jannis Uhlenhof, Dr. Thomas Handorf, and our Google Summer of Code students.

A number of people have provided very helpful comments at different stages of this manuscript. I would like to acknowledge the proofreading of Dr. Wolfram Liebermeister, Arturo Blázquez Navarro, Katharina Albers, Dr. Marcel Grunert, Dr. Natalie Stanford, Timo Lubitz, and Jannis Uhlenhof. I am sorry that I had so much to say.

Furthermore, I would like to thank the coordinators of the International Max Planck Research School for Computational Biology and Scientific Computing, Dr. Hannes Luz and Dr. Kirsten Kelleher. Your care about important things has made you indispensable for this work. Your care about the small things has made you unforgettable.

I would like to acknowledge funding from the International Max Planck Research School for Computational Biology and Scientific Computing, the Sonderforschungsbereich (SFB) 618, and the Marie Curie Early Stage Training. Further travel funds have been provided by FORSYS, the Gesellschaft für Biochemie und Molekularbiologie e.V., Google Inc., and BASF SE.

Finally, and most importantly I would like to thank my family and friends for providing moral support throughout the last years. Thanks for the encouragements, every pat on the back, and all the nights out to forget a day's setbacks. Without you all this work would not have been possible.

Contents

| | | |
|----------|----------------------------------------------------------|-----------|
| 1 | Introduction | 21 |
| 1.1 | Motivation and outline | 22 |
| 1.1.1 | Problems in current drug research | 22 |
| 1.1.2 | Improving drug admission rates | 23 |
| 1.1.3 | Contents of the introduction | 24 |
| 1.2 | Current state of drug research | 24 |
| 1.2.1 | Target-based approach | 24 |
| 1.2.2 | Phenotypic screening | 26 |
| 1.2.3 | Other approaches | 27 |
| 1.3 | Mathematical modelling of biological processes | 27 |
| 1.3.1 | What is Systems Biology? | 28 |
| 1.3.2 | Different levels of dynamic models | 29 |
| 1.3.3 | Examples of successful predictions | 30 |
| 1.4 | Drug target identification | 31 |
| 1.4.1 | Network approaches | 31 |
| 1.4.1.1 | Application to drug research | 32 |
| 1.4.1.2 | General network analyses | 32 |
| 1.4.2 | Stoichiometric approaches | 33 |
| 1.4.3 | Kinetic modelling approaches | 34 |
| 1.5 | Glycolysis in <i>Trypanosoma brucei</i> | 35 |
| 1.5.1 | General information on the disease | 35 |
| 1.5.2 | Available treatments | 35 |
| 1.5.3 | Potential new treatments | 36 |
| 1.5.4 | Available models of trypanosomal glycolysis | 37 |
| 1.5.5 | <i>In silico</i> determination of drug targets | 38 |
| 1.6 | The arachidonic acid pathway | 39 |
| 1.6.1 | Physiology | 39 |
| 1.6.1.1 | Pathway structure | 39 |
| 1.6.1.2 | Regulation | 39 |
| 1.6.1.3 | Downstream effects | 41 |

| | | |
|----------|-----------------------------------------------------------|-----------|
| 1.6.2 | Pathophysiology | 42 |
| 1.6.2.1 | Diseases involving the AA pathway | 42 |
| 1.6.2.2 | Drugs acting in the AA pathway | 42 |
| 1.6.3 | Modelling the AA pathway | 43 |
| 1.7 | Outline of this work | 44 |
| 1.7.1 | The big picture | 44 |
| 1.7.2 | Contents of this thesis | 45 |
| 2 | Methods | 47 |
| 2.1 | Modelling using ordinary differential equations | 47 |
| 2.1.1 | Problem formulation | 48 |
| 2.1.2 | Solutions in time | 48 |
| 2.1.3 | Structure of models | 49 |
| 2.1.3.1 | Stoichiometric matrix | 49 |
| 2.1.3.2 | Kinetics | 50 |
| 2.1.4 | Metabolic control analysis | 51 |
| 2.1.4.1 | Steady state | 51 |
| 2.1.4.2 | Elasticities | 51 |
| 2.1.4.3 | Control and response coefficients | 52 |
| 2.2 | Parameter estimation | 52 |
| 2.2.1 | Problem formulation | 52 |
| 2.2.2 | Algorithms and heuristics for optimisation | 53 |
| 2.2.3 | Statistical assessment of fit quality | 54 |
| 2.3 | Model representation | 54 |
| 2.3.1 | Systems Biology Markup Language | 54 |
| 2.3.2 | Model annotation | 55 |
| 2.3.3 | Model repositories | 56 |

I Finding and refining available biological knowledge **57**

| | | |
|----------|--------------------------------------------------------------|-----------|
| 3 | Semantic similarity measures for Systems Biology | 59 |
| 3.1 | Semantic information in Systems Biology | 60 |
| 3.1.1 | Using available information for modelling | 60 |
| 3.1.2 | Integrating ontologies | 65 |
| 3.1.3 | Similarity measure for entries in ontologies | 67 |
| 3.1.3.1 | Measures used to compare ontology entries | 67 |
| 3.1.3.2 | Biological applications for similarity measures | 69 |
| 3.1.4 | Principles of information retrieval | 70 |
| 3.1.5 | Standards for semantic information on the internet | 70 |

| | | |
|----------|-------------------------------------------------------------------------------|-----------|
| 3.2 | Comparing models and data sets based on semantic information | 71 |
| 3.2.1 | Combining ontologies | 71 |
| 3.2.1.1 | A library to integrate ontologies | 71 |
| 3.2.1.2 | Available frameworks incorporating semantic information | 72 |
| 3.2.1.3 | Database schema of the libSBAnnotation | 73 |
| 3.2.1.4 | Integrating information from different ontologies | 75 |
| 3.2.2 | Similarity measures for Biological Concepts | 75 |
| 3.2.2.1 | Li's modular similarity measure | 75 |
| 3.2.2.2 | A modular measure for comparing Biological Concepts | 79 |
| 3.2.3 | Similarity measures for annotated data sets and models | 79 |
| 3.2.3.1 | Comparing MIRIAM annotations | 79 |
| 3.2.3.2 | The preference-based measure | 80 |
| 3.2.3.3 | Vector space model | 81 |
| 3.3 | Retrieval, alignment, and clustering of models and data sets | 83 |
| 3.3.1 | Assessing the quality of different similarity measures | 83 |
| 3.3.2 | Evaluation of the different measures | 84 |
| 3.3.3 | Matching experimental data to models | 87 |
| 3.3.3.1 | Oscillating yeast genes | 87 |
| 3.3.3.2 | Arachidonic acid pathway | 88 |
| 3.3.4 | Matching Systems Biology models | 89 |
| 3.3.4.1 | Retrieving MAP kinase cascade models | 89 |
| 3.3.4.2 | Clustering of MAP kinase cascade models | 91 |
| 3.3.4.3 | Simple alignments of MAP kinase cascade models | 92 |
| 3.4 | Discussion | 92 |
| 3.4.1 | Retrieval of models and data sets | 92 |
| 3.4.2 | Criteria for model similarity | 94 |
| 3.4.3 | Quality of the presented measures | 95 |
| 3.4.4 | Limitations of the current method | 96 |
| 3.4.5 | Conclusion | 96 |
| 4 | Incorporating structural information into semantic similarity measures | 99 |
| 4.1 | Aligning biochemical networks | 100 |
| 4.1.1 | Comparing biochemical networks | 100 |
| 4.1.2 | General differences in network comparison algorithms | 101 |
| 4.1.3 | Merging Systems Biology models | 102 |
| 4.2 | Distributing semantic information in biochemical networks | 103 |

| | | |
|---------|-------------------------------------------------------------------------|-----|
| 4.2.1 | Feature propagation in biochemical networks | 105 |
| 4.2.1.1 | General idea behind feature propagation . . . | 105 |
| 4.2.1.2 | Mathematical formulation of feature propa- gation | 105 |
| 4.2.1.3 | Observations on feature propagation | 106 |
| 4.2.2 | Semantic propagation for merging network models . . . | 108 |
| 4.2.3 | Predicting missing annotations in biochemical networks | 109 |
| 4.2.4 | Implementation of propagation methods for SBML mod- els | 109 |
| 4.3 | Applications of improved model alignments | 111 |
| 4.3.1 | Improvements in the alignment of MAPK models . . . | 111 |
| 4.3.2 | Randomised removal of annotations and large scale analysis | 113 |
| 4.3.3 | Predicting annotations in a glycolysis model | 115 |
| 4.3.4 | Merging arachidonic acid pathways | 117 |
| 4.4 | Discussion | 118 |
| 4.4.1 | Combining structural and semantic information | 118 |
| 4.4.2 | Assessing the quality of the proposed measures | 119 |
| 4.4.3 | Comparison to existing approaches | 120 |
| 4.4.4 | Conclusion | 121 |

II Predicting possible drugs and drug targets 123

5 Identifying potential drug targets in ODE models 125

| | | |
|---------|------------------------------------------------------------------------------------------------|-----|
| 5.1 | Drug target identification by parameter optimisation | 126 |
| 5.1.1 | Drug target identification is a parameter estimation problem | 127 |
| 5.1.1.1 | Implications concerning solutions | 128 |
| 5.1.2 | Parameter identifiability | 129 |
| 5.1.2.1 | Why does identifiability matter? | 129 |
| 5.1.2.2 | Available approaches to find non-identifiable parameters | 130 |
| 5.1.3 | Host pathogen systems | 131 |
| 5.2 | Formalising the drug target identification problem | 131 |
| 5.2.1 | Setting up a parameter estimation to solve the drug target identification problem | 132 |
| 5.2.1.1 | Identifying and inserting kinetics | 132 |
| 5.2.1.2 | Affecting multiple targets with a single drug . | 134 |
| 5.2.1.3 | Preparing the objective function | 134 |

| | | |
|----------|------------------------------------------------------------------------------------------|------------|
| 5.2.2 | Different optimisation methods and what solutions to expect | 136 |
| 5.2.3 | Parameter identifiability | 138 |
| 5.2.4 | Host-pathogen interactions | 140 |
| 5.2.4.1 | Network selectivity in metabolic control analysis | 140 |
| 5.2.4.2 | Network selectivity in drug target identification | 141 |
| 5.3 | Results | 142 |
| 5.3.1 | Implementation | 142 |
| 5.3.2 | Linear pathway | 143 |
| 5.3.2.1 | Introduction of the model | 143 |
| 5.3.2.2 | Insights from the application of my framework | 143 |
| 5.3.3 | Glycolysis in <i>Trypanosoma brucei</i> | 145 |
| 5.3.3.1 | Prioritisation of drug targets by necessary effective inhibitor concentrations | 145 |
| 5.3.3.2 | What determines the quality of a target | 145 |
| 5.3.3.3 | Considering side effects in human cells | 146 |
| 5.3.4 | AA pathway | 148 |
| 5.3.4.1 | Preparation of model | 148 |
| 5.3.4.2 | Prior results | 149 |
| 5.3.4.3 | Results | 150 |
| 5.4 | Discussion | 154 |
| 5.4.1 | What has been achieved in this chapter | 154 |
| 5.4.2 | Comparison to other approaches | 154 |
| 5.4.2.1 | Predicting valuable drug targets | 154 |
| 5.4.2.2 | Identifiability | 157 |
| 5.4.3 | Biologically relevant results | 158 |
| 5.4.3.1 | Glycolysis in <i>Trypanosoma brucei</i> | 158 |
| 5.4.3.2 | Arachidonic acid pathway | 159 |
| 5.4.4 | Drug resistance through mutations | 161 |
| 5.4.5 | Outlook | 162 |
| 5.4.5.1 | Implementation | 162 |
| 5.4.5.2 | Tackling parameter uncertainties | 162 |
| 5.4.5.3 | Using non-identifiabilities to direct further research | 163 |
| 5.4.5.4 | Reverse drug target prediction enables mode-of-action identification | 164 |
| 6 | Drug-drug interactions | 165 |
| 6.1 | Introduction | 165 |
| 6.1.1 | Synergisms and antagonisms in current research | 165 |

| | | |
|----------|----------------------------------------------------------------------|------------|
| 6.1.1.1 | Synergistic drug combinations require lower doses | 166 |
| 6.1.1.2 | Antagonisms and resistance | 167 |
| 6.1.1.3 | What causes synergy? | 167 |
| 6.1.2 | Mathematical definitions of synergy | 168 |
| 6.1.2.1 | Null models of combined effects | 168 |
| 6.1.2.2 | How to detect synergy | 169 |
| 6.1.2.3 | How to quantify synergy | 170 |
| 6.2 | Methods | 171 |
| 6.2.1 | Synergism detection in ODE models | 171 |
| 6.2.2 | Choice of the null models | 171 |
| 6.2.3 | Computational detection of synergisms | 172 |
| 6.3 | Results | 174 |
| 6.3.1 | Glycolysis in <i>Trypanosoma brucei</i> | 174 |
| 6.3.1.1 | Detected synergisms | 174 |
| 6.3.1.2 | Detected antagonisms | 175 |
| 6.4 | Discussion | 177 |
| 6.4.1 | Synergisms in the trypanosomal glycolysis | 177 |
| 6.4.2 | Advantages of the employed method | 178 |
| 6.4.3 | Advantages of synergisms and antagonisms for drug research | 179 |
| 7 | Discussion | 181 |
| 7.1 | Achievements | 181 |
| 7.1.1 | A framework for drug target identification | 181 |
| 7.1.2 | Target predictions | 183 |
| 7.2 | Directing new research | 184 |
| 7.2.1 | How to proceed from predictions | 184 |
| 7.2.2 | Curing sleeping sickness | 184 |
| 7.2.3 | Reducing inflammatory responses | 185 |
| 7.3 | Extensions to my framework | 185 |
| 7.3.1 | Using web resources for drug prediction | 185 |
| 7.3.1.1 | Publicly available resources | 186 |
| 7.3.1.2 | Identifying drugs for selected targets | 186 |
| 7.3.2 | Considering parameter uncertainties | 187 |
| 7.4 | Systems Biology for drug research | 188 |
| 7.4.1 | Comparing different approaches | 188 |
| 7.4.1.1 | Predictions based on different information | 188 |
| 7.4.1.2 | Advantages of ODE models | 188 |
| 7.4.2 | Limitations of predicted drugs | 189 |
| 7.4.2.1 | Potential lack of efficacy | 189 |

| | | |
|----------|------------------------------------------------------------------------------------------------------------------------|------------|
| 7.4.2.2 | Potential side effects | 190 |
| 7.4.2.3 | Potential ADME problems | 190 |
| 7.4.2.4 | Link to personal medicine | 191 |
| A | Supplementary Figures and Tables | 235 |
| A.1 | Parameters and data | 235 |
| B | Supplementary methods | 241 |
| B.1 | Mathematical details on similarity measures | 241 |
| B.1.1 | Statistical significance of retrieved models | 241 |
| B.1.1.1 | Null model for the VSM | 241 |
| B.1.1.2 | Bayesian estimation of a p-value | 242 |
| B.1.1.3 | Analytic derivation of a p-value | 242 |
| B.2 | Mathematical details on drug target identification | 245 |
| B.2.1 | Different inhibition kinetics | 245 |
| B.2.1.1 | Irreversible Michaelis-Menten kinetics | 245 |
| B.2.1.2 | Reversible Michaelis-Menten kinetics | 248 |
| B.2.1.3 | General considerations on inhibition/activation kinetics | 250 |
| B.2.2 | Construction of a objective function for drug target identification | 251 |
| B.2.2.1 | Proof: All objectives are fulfilled if the objec- tive function has a value smaller than one | 251 |
| B.2.3 | Parameter identifiability | 252 |
| B.2.3.1 | Mathematical reasoning for investigating χ^2 | 252 |
| B.2.3.2 | Rules for interchangeabilities | 254 |
| B.2.3.3 | Algorithm for identifying drug interchange- abilities | 255 |
| B.2.4 | Extending my definition of network selectivity | 256 |
| B.2.4.1 | Selectivity in higher order drug treatments | 256 |
| B.2.4.2 | Selectivity in between multiple models | 257 |
| B.2.5 | A mathematical model to relate necessary drug con- centrations to probabilities of resistance development | 257 |
| B.3 | Mathematical details of synergism analysis | 260 |
| B.3.1 | Rewriting Bliss drug interaction models | 260 |
| C | Models and objective functions for the target identification | 263 |
| C.1 | Linear chain | 263 |
| C.1.1 | Potency of different inhibition types | 263 |
| C.1.2 | Setting up the objective function | 265 |
| C.1.3 | Results | 267 |

| | | |
|-------|-------------------------------------------------------------|-----|
| C.2 | Glycolysis in <i>Trypanosoma brucei</i> | 272 |
| C.2.1 | Setting up the objective function | 272 |
| C.2.2 | Results | 272 |
| C.3 | Glycolysis in human erythrocytes | 276 |
| C.3.1 | Setting up the objective function | 277 |
| C.3.2 | Results | 277 |
| C.4 | Arachidonic acid pathway in different human cells | 280 |
| C.4.1 | Setting up the objective function | 282 |
| C.4.2 | Results | 283 |

List of Abbreviations

| | |
|-------|----------------------------------------|
| AA | arachidonic acid |
| ADP | adenosine diphosphate |
| AIC | Akaike information criterion |
| ALD | fructose bisphosphate aldolase |
| ATP | adenosine triphosphate |
| AUC | area under the curve |
| cAMP | cyclic adenosine monophosphate |
| COX | cyclooxygenase |
| CV | controlled vocabulary |
| DHAP | dihydroxyacetone phosphate |
| DNA | deoxyribonucleic acid |
| EC | enzyme classification |
| EET | epoxyeicosatrienoic acids |
| ENO | enolase |
| G3P | glycerol-3-phosphate |
| GAPDH | glyceraldehyde phosphate dehydrogenase |
| GPDH | glycerol-3-phosphate dehydrogenase |
| GPO | glycerol-3-phosphate oxidase |
| HETE | hydroxyeicosatetraenoic acid |

HK hexokinase
IVP initial value problem
LOX lipoxygenase
MCA metabolic control analysis
mPGES1 microsomal PGE₂ synthase-1
NSAID non-steroidal anti-inflammatory drug
ODE ordinary differential equation
PDB Protein Data Bank
PGI phosphoglucoisomerase
PGK phosphoglycerate kinase
PGM phosphoglycerate mutase
PHGPx phospholipid hydroperoxide glutathione peroxidase
PPAR peroxisomal proliferator-activated receptor
PPI protein-protein interaction
PT pyruvate transporter
RNA ribonucleic acid
SBML Systems Biology Markup Language
SBO Systems Biology Ontology
SCE small chemical entity
SHAM salicylhydroxamic acid
TAO trypanosomal alternative oxidase
THT trypanosomal hexose transporter
TPI triosephosphateisomerase
TVSM topic-based vector space model
TX thromboxane

UTR untranslated region

VSM vector space model

List of Figures

| | | |
|-----|-------------------------------------------------------------------|-----|
| 1.1 | The arachidonic acid pathway and its regulation | 40 |
| 1.2 | Workflow overview | 45 |
| 2.1 | Small example network | 49 |
| 3.1 | Workflow overview | 61 |
| 3.2 | Example of WordNet structure | 67 |
| 3.3 | libSBAnnotation database scheme | 74 |
| 3.4 | Structure of Biological Concepts within model collections . . . | 76 |
| 3.5 | Explanation of preference-based similarity measure | 81 |
| 3.6 | Model retrieval from data | 88 |
| 3.7 | Model retrieval of MAP kinase cascades | 90 |
| 3.8 | Model clustering | 91 |
| 3.9 | Model alignment | 93 |
| 4.1 | Workflow overview | 102 |
| 4.2 | Equivalent reaction networks | 104 |
| 4.3 | Information propagation in a network | 106 |
| 4.4 | Logical dependencies in between SBML elements | 110 |
| 4.5 | Detailed model alignment | 112 |
| 4.6 | Assessment of alignment quality | 114 |
| 4.7 | Assessment of quality of annotation prediction | 116 |
| 5.1 | Workflow overview | 126 |
| 5.2 | Objective function landscapes | 128 |
| 5.3 | Different objectives for time course data | 135 |
| 5.4 | Processes contributing to the effectivity of an administered drug | 141 |
| 6.1 | Cooperativity of inhibitors of GPO and GK | 177 |
| 7.1 | Workflow overview | 182 |
| C.1 | Structure of linear example model | 263 |

| | | |
|------|----------------------------------------------------------------------------|-----|
| C.2 | Inhibitions in the linear chain 1 | 267 |
| C.3 | Inhibitions in the linear chain 2 | 269 |
| C.4 | Inhibitions in the linear chain 3 | 270 |
| C.5 | Inhibitions in the linear chain 4 | 271 |
| C.6 | Network structure of Albert's glycolysis model | 273 |
| C.7 | Network structure of Achcar's glycolysis model | 274 |
| C.8 | Retrieval of glycolysis models | 276 |
| C.9 | Network structure of Holzhütter's erythrocyte metabolism model | 277 |
| C.10 | Structure of Yang's arachidonic acid pathway model | 280 |
| C.11 | Targets and their effects in the AA pathway | 285 |
| C.12 | Effects of single inhibitors in AA pathway | 286 |
| C.13 | Effects of inhibitors in combination with a PLA ₂ inhibitor . . | 287 |
| C.14 | Nonidentifiabilities in the AA pathway 1 | 288 |
| C.15 | Nonidentifiabilities in the AA pathway 2 | 289 |

List of Tables

| | | |
|-----|--------------------------------------------------------------------------------|-----|
| 1.1 | List of currently applied trypanocidal drugs. | 36 |
| 1.2 | Trypanocidal drugs and their targets | 37 |
| 1.3 | Drugs and their targets in the arachidonic acid pathway | 44 |
| 3.1 | List of biologically relevant web repositories | 62 |
| 3.2 | Web repositories included in libSBAnnotation | 73 |
| 3.3 | Small model group example | 83 |
| 3.4 | Evaluation of similarity measures | 85 |
| 5.1 | Network selectivities of glycolytic enzymes | 147 |
| 5.2 | Sequence identity of glycolytic proteins | 156 |
| 6.1 | Synergisms of targets in trypanosomal glycolysis | 174 |
| 6.2 | Antagonisms of targets in trypanosomal glycolysis | 175 |
| 6.3 | Predictivity of synergism models | 176 |
| A.1 | f_{rts} values | 236 |
| A.2 | f_{qsm} values | 237 |
| A.3 | Large model group example | 238 |
| A.4 | Sensity of the model retrieval to f_{rts} | 239 |
| B.1 | Inhibition factors in Michaelis-Menten type kinetics | 250 |
| C.1 | Necessary inhibitor concentrations in trypanosomal glycolysis . | 275 |
| C.2 | Tolerable inhibitor concentrations in erythrocyte glycolysis . . | 278 |
| C.3 | Network selectivity in glycolysis | 279 |
| C.4 | Experimental data for the arachidonic acid pathway | 281 |
| C.5 | Necessary inhibitor concentrations reducing leukotriene production | 283 |
| C.6 | Necessary inhibitor concentrations reducing prostaglandin production | 284 |
| C.7 | Possible drug combinations in AA pathway | 290 |

Chapter 1

Introduction

Contents

| | | |
|------------|-------------------------------------------------------|-----------|
| 1.1 | Motivation and outline | 22 |
| 1.1.1 | Problems in current drug research | 22 |
| 1.1.2 | Improving drug admission rates | 23 |
| 1.1.3 | Contents of the introduction | 24 |
| 1.2 | Current state of drug research | 24 |
| 1.2.1 | Target-based approach | 24 |
| 1.2.2 | Phenotypic screening | 26 |
| 1.2.3 | Other approaches | 27 |
| 1.3 | Mathematical modelling of biological processes | 27 |
| 1.3.1 | What is Systems Biology? | 28 |
| 1.3.2 | Different levels of dynamic models | 29 |
| 1.3.3 | Examples of successful predictions | 30 |
| 1.4 | Drug target identification | 31 |
| 1.4.1 | Network approaches | 31 |
| 1.4.2 | Stoichiometric approaches | 33 |
| 1.4.3 | Kinetic modelling approaches | 34 |
| 1.5 | Glycolysis in <i>Trypanosoma brucei</i> | 35 |
| 1.5.1 | General information on the disease | 35 |
| 1.5.2 | Available treatments | 35 |
| 1.5.3 | Potential new treatments | 36 |
| 1.5.4 | Available models of trypanosomal glycolysis | 37 |

1.1. MOTIVATION AND OUTLINE

| | | |
|------------|----------------------------------------------------------|-----------|
| 1.5.5 | <i>In silico</i> determination of drug targets | 38 |
| 1.6 | The arachidonic acid pathway | 39 |
| 1.6.1 | Physiology | 39 |
| 1.6.2 | Pathophysiology | 42 |
| 1.6.3 | Modelling the AA pathway | 43 |
| 1.7 | Outline of this work | 44 |
| 1.7.1 | The big picture | 44 |
| 1.7.2 | Contents of this thesis | 45 |

1.1 Motivation and outline

1.1.1 Problems in current drug research

Decrease in research productivity Over the last years, the rate at which newly developed drugs have been approved continuously declined [Mathieu, 2007]. Amongst the few newly approved drugs an increasing number belongs to the group of follow-ups. These drugs target the same proteins as other approved drugs and therefore do not treat new diseases. In most cases they just provide a slight benefit over an existing treatment and replace a drug whose patent protection is running out. Even though these drugs provide little to no value to society, they are responsible for most of the revenue of the big pharma companies [Booth and Zimmel, 2003]. While the development of follow-ups of blockbuster drugs (leading to more than 1 billion dollar of revenue per year) is financially advantageous for companies, going after novel targets and curing new diseases bears lots of risks that can delay the drug development process [Sams-Dodd, 2005]. Highly problematic amongst such drug candidates are failures because of lack of *in vivo* efficacy and toxicity as they are comparably frequent and only detected in late stages of the development in which they have already lots of financial resources [Kola and Landis, 2004]. Thus, for pharma companies there is less commercial advantage in pursuing the development of drugs against novel targets [Ma and Zimmel, 2002].

Increase in development costs Due to the decline in output of new drugs the drug development process is becoming more and more expensive. It has been estimated that the average drug costs in between 800 million [DiMasi et al., 2003] and 1 billion dollars [Adams and Brantner, 2006]. Although these numbers are highly quoted, they may overestimate the real costs by up

1.1. MOTIVATION AND OUTLINE

to one order of magnitude [Light and Warburton, 2011]. For example half of the estimated costs stem from projected interest, that might have been achieved if the development costs would have been invested on the stock market. From the money actually spent on research and development only around 20% are spent on candidates which will become drugs [DiMasi et al., 2003]. The remaining 80% are spent on unsuccessful drug candidates, which turn into stronger financial damages the later they fail [Ashburn and Thor, 2004].

Apart from the financial burden on the health care system, the high costs of drug development have another aggravating consequence: they reduce the applicability of commercial pharmaceutical research to certain diseases. As the development of a drug has to be paid for by its beneficiaries, there is a lack of commercial interest in rare diseases or diseases only occurring in Third World Countries. Therefore, drug development for these diseases has to be taken over by non-profit academic research [Trouiller et al., 2002].

1.1.2 Improving drug admission rates

Selection of efficacious and safe targets The target-based approach for drug discovery follows two basic steps. First, targets are selected on the basis of biological knowledge, and second, drugs are designed to specifically inhibit these targets. If the initial selection of a target is not optimal, problems of drug candidates such as lack of efficacy or toxicity will arise in trials. Therefore, much more effort should be put into the target selection, e.g. by applying a combination of theoretical considerations and practical experiments to identify effective and safe targets. Potential candidates against ineffective targets will thus be sorted out as early as possible reducing total development costs.

Systems Biology in drug target identification For the purpose of identifying efficacious and non-toxic target, concepts and methods of Systems Biology can be applied to the target identification problem. In principle, some of these methods are already used in drug development and it is consequential to use them to tackle the selection of good targets [Butcher et al., 2004]. Methods for the identification of targets include the construction of mathematical models of relevant biochemical processes and the *in silico* simulation of the effects of potential drugs on them. Some of these predictions can afterwards be verified in experiments, supporting the selection of targets.

If drug research pursues this Systems Biology approach for target identification, the modulation of these targets has already been shown to be effective and safe. Thus, drugs against them will be less likely to fail in later

1.2. CURRENT STATE OF DRUG RESEARCH

stages of the development process. The inclusion of these methods will ultimately reduce the costs of drug development. Therefore, research on some diseases, that have been neglected for commercial reasons, may again become interesting to pharma companies.

1.1.3 Contents of the introduction

Within this thesis I tackle the problem of drug target identification with concepts from Systems Biology. Before I go into detail about my work, I will give an introduction to the field. The introduction will cover current paradigms in drug research and development. I will highlight some basic concepts of Systems Biology and list different approaches of how biological models have been used in drug research. Finally, two example systems are introduced: the glycolysis in *Trypanosoma brucei*, which is the pathogen causing the African sleeping sickness, and the arachidonic acid pathway in humans, in which inflammatory mediators are produced. These systems will be investigated throughout this work.

1.2 Current state of drug research

In the last decade drugs have mainly been developed following two approaches, (i) the target-based approach or (ii) the phenotypic screening [Swinney and Anthony, 2011]. I will discuss these different approaches in the following section.

1.2.1 Target-based approach

Magic bullets Rational development of new drugs has been driven by the idea that only one single protein needs to be targeted to cure a certain disease and that each such protein can be successfully targeted by a single drug [Lindsay, 2003]. Historically this is motivated by the work of Paul Ehrlich at the end of the 19th century [Keith et al., 2005]. He discovered that he could synthesise a dye which would specifically colour a certain pathogen. The combination of such a selective molecule with a toxin lead to the idea of the “magic bullet”, a drug which would specifically kill a certain organism. Paul Ehrlich later managed to synthesise the first magic bullet, the drug arshenammine, which kills *Treponema pallidum*, the pathogen causing syphilis [Ehrlich, 1913].

1.2. CURRENT STATE OF DRUG RESEARCH

How does the target-based approach work? Following this idea drug research is first identifying a target that is relevant to the disease, and second, effective and safe drugs are designed against this target. The whole process takes five to ten years and is divided into six consecutive steps [Bleicher et al., 2003, Sams-Dodd, 2005].

- Potential targets are selected based on biological knowledge or genomic data.
- Targets are experimentally validated, e.g. by investigating knock-outs or transgenic animals.
- An assay is developed in which compounds can be tested for their activity against the selected target.
- A library of compounds is screened using this assay to find initial “hits” [Davis et al., 2005].
- Hit compounds are further optimised for binding to the target [Alanine et al., 2003].
- Drug-like properties of the resulting “lead” molecules are further optimised, which will result in the final drugs that can afterwards be tested in clinical trials [Lipinski, 2004].

Advantages and drawbacks The target-based approach to drug discovery has been pursued by the pharma industry because it is a very directed and rational process. In each of the steps very clear requirements to the target or the drug can be formulated, which is advantageous for large companies to measure the success of a development process [Sams-Dodd, 2005]. Furthermore, the rational approach can be applied to many problems as it has no explicit requirements on the investigated biological system, except for initial knowledge on the mechanisms behind the disease of interest. However, its major disadvantage is its declining productivity regarding drugs with a novel mode-of-action [Mathieu, 2007]. One possible cause of this lack of success is the suboptimal initial selection of targets. This initial selection should focus more on the efficacy and safety of a drug targeting it to reduce attrition rates in later stages [Paul et al., 2010].

Single drug treatments One important idea behind the target-based approach is the magic bullet, which is the conception that for every disease there exists a single drug acting on a single target being able to cure the illness. For many diseases such magic bullets do exist. A very obvious example

1.2. CURRENT STATE OF DRUG RESEARCH

of such an illnesses is diabetes mellitus, which is caused by the inability to produce sufficient amounts of insulin or an insensitivity of its receptor. This disease can simply be treated by injecting insulin intravenously on a regular basis [Banting and Best, 1922].

Multi drug treatments Regardless of the many available examples, the assumption that diseases can generally be treated with a single drug might not always be correct. On the one hand, even if a disease is caused by a single factor it might become necessary to treat it using multiple drugs for various reasons. First, as robustness is a central property of biological networks, attacking it at one point might not be enough to continuously achieve a certain result [Albert et al., 2000, Csermely et al., 2005, Hopkins, 2008]. Second, disturbing many targets in parallel with various drugs might give the same effect at much lower individual drug doses than a single drug treatment [Korcsmáros et al., 2007]. This can become necessary if the high dose of the single drug leads to severe side-effects [Farr and Bacon, 1995]. Furthermore, the combination of multiple drugs can increase the selectivity of a treatment [Lehár et al., 2009] or it can reduce the rate of drug resistance development [Michel et al., 2008, Yeh et al., 2009]. Finally, if two approved drugs are combined, the resulting treatment has the advantage that it can enter the market comparably fast [Borisy et al., 2003].

On the other hand, many investigated diseases can have multiple causes. This includes diseases that can be caused by different mutations on the same protein. If the disease is to be treated with a drug against this protein, each mutation can in principle require a different drug. Therefore, a universally applicable treatment should involve all of these drugs [Radhakrishnan and Tidor, 2008]. In addition, it includes diseases requiring several successive mutations, such as cancer, which has been estimated to be caused by 4-7 independent mutations in most of the cases [Balmain et al., 1993]. Such multicausal diseases can not be expected to be successfully treated with a magic bullet. Thus, these cases require the selection of multiple targets which should be thoroughly chosen based on available biological information [Csermely et al., 2005]. For examples of treatments involving multiple drugs the reader is referred to [Zimmermann et al., 2007, Hopkins, 2008].

1.2.2 Phenotypic screening

Compared to the target-based approach phenotypic screening works the other way around. First, a biological example system for a disease is developed. These example systems can vary heavily in size ranging from single cells [Yarrow et al., 2003] to complete organisms [Gehrmann et al., 2000]. With

1.3. MATHEMATICAL MODELLING OF BIOLOGICAL PROCESSES

the help of these systems large compound libraries are scanned for substances with an *in vitro/vivo* activity. The result of such a high throughput scan is a selection of efficacious compounds for which the mode-of-action is unknown. This mode-of-action then has to be elucidated in order to further optimise drug-like properties of the compound and to be able to assess its toxicity.

Given a relevant biological example system and a large compound library the phenotypic approach is comparably easy to follow [Borisy et al., 2003, Yeh et al., 2006, Apsel et al., 2008, Sharlow et al., 2009]. However, the required biological system might not always be available, either because the disease is not understood well enough, the affected system cannot be tested in an ethically sound manner, or simply because large scale tests on this system are too expensive.

In principle the target-based approach is preferred by the pharma industry. However, most of the newly approved drugs with a new mode of action have been found by phenotypic screens [Swinney and Anthony, 2011]. Therefore, it has been argued that biological assays are more likely to produce drugs which are effective and safe *in vivo* [Butcher, 2005].

1.2.3 Other approaches

Apart from these two main approaches to drug research, several other ways to obtain successful treatments exist. First, drugs can be based on naturally occurring substances. An example of such a drug is fondaparinux, a cleavage product of heparin that binds to antithrombin III to inhibit Factor Xa thus inhibiting blood coagulation [Choay et al., 1983]. Second, treatments like vaccines and antibodies can directly be derived from natural processes. It should be noted that in the strict sense these “biologics” can be regarded as being target-based [Swinney and Anthony, 2011]. Furthermore, drugs can be repositioned when one of their side-effects can be used to treat another disease. Examples of such a drug repositioning are Viagra, which has been observed to cure erectile dysfunctions in clinical trials, and Comtan, which is primarily used to treat Parkinson’s disease and has been found to be useful against tuberculosis [Kinnings et al., 2009].

1.3 Mathematical modelling of biological processes

The desire to simplify the complex reality around us is an essential part of human nature. Our ability to abstract the world influences the way in which we perceive it and enables us to judge the consequences of potential actions

1.3. MATHEMATICAL MODELLING OF BIOLOGICAL PROCESSES

in advance. Therefore, models, which are our abstract interpretations of processes occurring around us, are the main drivers of human behaviour and major contributors to our evolutionary advancement.

Models shape our understanding of the world in all stages of our life. In our early years for example we observe that whenever we drop things they fall down. From a large number of these observations we deduce the general principle that all objects tend to move downwards if they are not stopped by something else. Having this general principle in mind, we get along quite well with our lives. However, this view is challenged when we learn that the earth is a sphere and that objects on the opposite side of it do not fall upwards. At this point we have to move to more complex explanations of our empirical findings: First, to the idea that everything moves towards the centre of the earth, and then, when we learn about the solar system, to the idea of gravitation.

1.3.1 What is Systems Biology?

Systems Biology can be perceived as different things: a scientific area, a collection of methodologies, or a philosophy of how the scientific process can acquire knowledge. Regardless of what Systems Biology is regarded to be, it has been characterised by the same properties:

- Systems Biology is interdisciplinary. It combines knowledge and methods from physics, chemistry, biology, mathematics, philosophy, and computer science, bridging the gap in between these disciplines.
- Systems Biology is integrative. It uses diverse kinds of data ranging from physical properties of single molecules to the behaviour of populations of complete organisms and it is able to incorporate huge amounts of information stemming from various “omics” measurements, such as transcriptomics and proteomics.
- Systems Biology is holistic. It integrates data to deduce the “big picture” rather than processing information individually.
- Systems Biology is structured. Its results are presented in the form of testable hypotheses, such as mathematical models of biological processes. These testable hypothesis allow for the progression of research through the so-called cycle of Systems Biology. This cycle denotes the idea that a hypothesis can be tested in experiments, which might falsify it and call for the integration of this new information into a new, testable hypothesis.

1.3. MATHEMATICAL MODELLING OF BIOLOGICAL PROCESSES

As a consequence of the approach of Systems Biology we can never expect a model, which describes a certain biological process, to be right. We can only say that a model is fit to represent a particular behaviour and make proper predictions under certain conditions. Instead of being verified, models can easily be falsified by contradicting observations, which ultimately reduces the number of possible explanations for a biological phenomenon [Popper, 1934]. This fact should always be kept in mind when dealing with models. However, a model is can still be very useful to describe the current knowledge about a particular system and as such a comprehensive description it has the potential to replace databases as our current resources of information [Aldridge et al., 2006].

For the purpose of knowledge integration, Systems Biology research progresses in different manners: top-down, in which one starts from observations and tries to explain them by increasingly complex models, and bottom-up, where networks are built from knowledge about molecular interactions and refined with the help of experimental data [Bruggeman and Westerhoff, 2007]. For larger applications, neither of the two approaches can be followed strictly because for example the description of a complete organism by the interactions of individual atoms is currently by far too complex. One way to circumvent this problem is to use a layered design, describing an individual by its organs, which is described as a network of interacting cells, which are described as reaction networks, and so on. For this principle the term middle-out has been coined [Brenner, 2010].

1.3.2 Different levels of dynamic models

Depending on the amount of knowledge which is available on a certain system, the amount and detailedness of available experimental data, and the specific scientific question that should be answered with it, models should describe different aspects of a system and therefore use different mathematical formalisms. As not all of these allow for the prioritisation of drug targets, I will give a list of some formalisms and their application areas in the following.

Boolean models For modelling large gene regulatory networks Boolean models are often the most appropriate formalism. In such models the variables describe the state of a gene to be either on or off, or expressed and not expressed, respectively, and the change in variable values over discrete time steps is described by Boolean update rules.

1.3. MATHEMATICAL MODELLING OF BIOLOGICAL PROCESSES

Discrete models In discrete models the variables can assume more than two states. This allows for the description of models in which genes can have multiple effects depending on the degree to which they are expressed. Update rules thus do include thresholds above which a certain gene regulation becomes active.

Ordinary differential equation models ODE models describe variables and the time by continuous values and describe the time evolution of a system by the rates in which the variables change over time. These kind of models have been applied to various kinds of systems including gene regulation, metabolic reactions, and signalling cascades.

Stochastic models Whenever systems involving few molecules of different substances are described, stochastic models can be used to trace the reactions of individual molecules. This is especially useful to observe the effects of random fluctuations on the behaviour of a model.

Spatial models ODE and stochastic models can be further extended to describe the localisation of molecules within a compartment. Such models are for example used when diffusion of substances has a high impact on the dynamics of the system or when the spatial distribution of molecules is important for further considerations.

Amongst these different formalism, ODE models are the most simple approach which allows for the inclusion of inhibitors as continuous variables. Using this formalism it is therefore possible to predict the quantitative relation between the concentration of a drug and its effect on the treated organism, which can be a desirable outcome for my applications.

1.3.3 Examples of successful predictions

The quality of a mathematical model is usually judged by its predictive power. Because of that two examples of successful predictions should be mentioned, which underline that the concept of modelling has high implications in biology. Probably the most well known model is the Hodgkin-Huxley model of action potential generation and transmission in squid axons [Hodgkin and Huxley, 1952]. As a conclusion to more than a decade of work the model is able to explain the behaviour of a cell from the action of ion channels, it proposed a general formula describing the actions of ion channels, and provided a general framework in which scientific research would be performed afterwards [Hausser, 2000].

1.4. DRUG TARGET IDENTIFICATION

Apart from their ability to describe observed biological results, bottom-up models can also be used to predict general principles of how pathways behave under certain conditions. An example of a successful prediction can be found in [Klipp et al., 2002]. In their article Klipp *et al.* investigated how and in which temporal order a limited amount of total protein should catalyse different reactions in a linear chain in order to make the reaction chain maximally effective. The theoretical results have afterwards been reproduced experimentally for the amino acid metabolism of *E. coli.*, and they supported the idea that within a linear pathway the expression of functionally successive enzymes is delayed and enzymes appearing more early in the chain have to be expressed more strongly [Zaslaver et al., 2004].

1.4 Drug target identification

Over the course of the last years the problem of identifying drug targets with the help of Systems Biology has been subject to extensive research (e.g. [Singh and Ghosh, 2006, Yang et al., 2008, Schoeberl et al., 2009]). Various approaches have been published, which differ not only in algorithmic details but also in the type of information used. Depending on the amount of knowledge that is available on a certain studied organism or pathway and the detailedness of the scientific questions asked, different types of models are used. Information in the format of a graph can be used as well as stoichiometric or kinetic models. As a result of the increasing amount of information that is evaluated along the different approaches, kinetic models will lead to more precise predictions as approaches based solely on network data.

1.4.1 Network approaches

Approaches based on data in the form of graphs have been applied in various areas of drug research, e.g. in the identification of potent drug targets or in the investigation of a drug's mode of action [Ágoston et al., 2005, Iorio et al., 2009]. The biological meaning of these graphs, however, can be completely different. Nodes can represent drugs, targeted proteins, diseases, or the gene involved in them while edges can indicate binding, influences, causal relations, or different kinds of similarities or commonalities. Various kinds of networks involving information on drugs have been compiled, and they have given insights into properties of successful drugs and the processes underlying their development. A small selection of different kinds of network analyses should be presented in the following.

1.4. DRUG TARGET IDENTIFICATION

1.4.1.1 Application to drug research

Finding potential drug targets The most important application of networks in pharmaceutical research is the identification of potent drug targets. In their work, Wu *et al.* [Wu et al., 2010] have used a molecular interaction network in combination with gene expression data after single drug treatment to predict drug-affected subnetworks. This is formulated as a maximum weight subgraph problem, which also incorporates the effects of multiple inhibitions and takes care of potential side-effects of the treatment. In another example potential treatments in signalling networks are investigated for their effects under diseased and healthy conditions. Ruths *et al.* [Ruths et al., 2006] show that the problem of achieving this with the minimal number of drugs is NP-hard and provide a heuristic to solve it. A completely different kind of network has been investigated by Vazquez *et al.* In their work they try to find a minimal set of known treatments, which will completely eradicate a population of partially resistant pathogens [Vazquez, 2009]. Tools using network topology to predict drug targets and further publications including network data for specific diseases can be found in a review by Berger & Iyengar [Berger and Iyengar, 2009].

Identifying drug effects Apart from the identification of new drug targets network based approaches can also be used to investigate unknown targets for existing drugs. Most of the available methods exploit networks of drugs which are based on their similarity and are used to infer modes-of-action by the targets of similar drugs. This is based on the observation that if two drugs bind mostly to the same proteins, their effects and side-effects will be similar [Fliri et al., 2005]. The drug similarity networks have been built from different kinds of data like gene expression data after treatment [Xing and Gardner, 2006, Iorio et al., 2009] or molecule structure and side-effect similarity [Campillos et al., 2008]. Information on novel targets for a certain drug can also be used to find novel applications for it. This is for example done by the web resource PROMISCUOUS, which combines drug-target and protein-protein relations that can be used to find new drug applications [von Eichborn et al., 2011]. In general, it should be noted that one is more likely to find new targets if the investigated drugs are small molecules [Hopkins et al., 2006].

1.4.1.2 General network analyses

Network approaches cannot only be used to answer detailed questions relevant to certain diseases but they can also be analysed in a general way.

1.4. DRUG TARGET IDENTIFICATION

Graphs compiled from drugs, their targets, and the interactions between those has been used to find general properties of drug-target networks, like the finding that most drugs are non-essential hubs in protein-protein interaction (PPI) networks [Yildirim et al., 2007, Ma’ayan et al., 2007]. In gene-disease networks it has been shown that interacting proteins have a higher chance to be involved in the same disease [Goh et al., 2007], which is also supported by other investigations [Luo et al., 2007]. Furthermore, networks of approved drugs and their application areas have also been constructed and investigated [Nacher and Schwartz, 2008].

1.4.2 Stoichiometric approaches

Recent developments in genomic sequence analysis have led to a rapid increase in the amount of genomic data available for various organisms. Using this sequence data, the reconstruction of the complete metabolic networks of various organisms has become feasible. Although these networks cannot be guaranteed to be complete, they are integrated resources of the current knowledge on a certain organism and some of them have already provided promising preliminary results (e.g. [Herrgård et al., 2008, Yus et al., 2009]).

In principle, stoichiometric networks do not define any dynamical properties of the described system. However, using the assumption that a metabolic network in a living organism will always try to operate close to a steady state, one can make predictions on the metabolic fluxes through the reactions [Heinrich and Rapoport, 1974]. This steady state assumption is usually justified with the fact that no metabolite can endlessly be consumed or produced. Given this assumption one can make some predictions on the behaviour of an investigated network as for example the ability of a knock-out mutation to survive. Based on this idea a number of drug target identification approaches have been developed.

Yeh *et al.* [Yeh et al., 2004] constructed a metabolic network for the Malaria causing pathogen *Plasmodium falciparum*. Using this network they have revealed that most of the known drug targets are so-called “chokepoints”. A chokepoint reaction is defined as either the only reaction producing a certain metabolite or the only one consuming it. This kind of analysis has also been used to investigate the metabolism of *Entamoeba histolytica* [Singh et al., 2007b]. In order to reduce the number of predicted targets and to yield safer targets, genomic data of the human can be incorporated to target only parasitic enzymes which have no orthologue. Another idea to reduce side effects of a treatment is to reduce the number of affected, non-disease-related compounds [Sridhar et al., 2006]. Furthermore, as stoichiometric approaches lead to large result sets, it is possible to prioritise

1.4. DRUG TARGET IDENTIFICATION

targets by various features as druggability, their expression, or their phylogenetic distribution [Crowther et al., 2010].

Folger *et al.* [Folger et al., 2011] have constructed a genome scale metabolic model of cancer cell lines in NCI-60 (a set of 59 cell lines). Using flux balance analysis [Edwards et al., 2001] they predicted the effects of knock-downs of all reactions on the speed of growth, i.e. the biomass production. To address potential side-effects the same knock-downs were also simulated in a full scale human model [Duarte et al., 2007] using biomass and internal energy production in terms of adenosine triphosphate (ATP) as objective functions. Afterwards the selectivity of the targets for NCI-60 instead of non cancer cells has been calculated. The simulation of single targets resulted in few selective treatments, however their number can be increased when considering synergistic dual knock-downs. This is in accordance with experimental results showing higher specificity for multi-target treatments [Lehár et al., 2009].

1.4.3 Kinetic modelling approaches

Although drug target identification approaches based on stoichiometric models have been successful in a number of cases, their results include comparably large numbers of possible targets between which no further distinction is possible and some questions cannot be answered with them at all. A biological question requiring a more detailed kinetic model has been posed by Stites *et al.* [Stites et al., 2007]. In their work the authors ask which conformation of the Ras protein should be targeted to achieve a higher signalling inhibition in cancer than in wild type cells.

Dynamic models of signalling processes have also already been used by the industry. The company Merrimack Pharmaceutical has focused on the development of drugs against epidermal growth factor receptors in the treatment of cancer and has used various models to evaluate their importance (e.g. [Schoeberl et al., 2009]).

Methodologically one can distinguish two different types of analysis that can be used for the purpose of drug target identification: Steady state and dynamical analysis. In steady state analysis the influence of infinitesimal changes of some variables on the long term behaviour of others is investigated. The results of these analyses are used to find reactions that exert a large control over certain aspects of the network (e.g. [Bakker et al., 2000b, Hornberg et al., 2005a, Murabito et al., 2011]). Dynamic analysis can not only incorporate changes in the steady state behaviour of variables but also their detailed dynamic behaviour [Tveito and Lines, 2009]. Furthermore, it is able to simulate the effects of treatments at effective drug concentrations.

1.5. GLYCOLYSIS IN *TRYPANOSOMA BRUCEI*

For the purpose of identifying potent targets in kinetic models, different methods and tools have been developed. Steady state analysis can be performed using various tools that are capable of performing Metabolic Control Analysis such as CoPaSi [Hoops et al., 2006]. Dynamic analysis has been performed using various methods which deal with the combinatorial problem of multiple target interventions in a different way [Araujo et al., 2005, Dasika et al., 2006, Yang et al., 2008, Tveito and Lines, 2009], tools for their automatic analysis, however, are scarce [Schulz et al., 2009].

1.5 Glycolysis in *Trypanosoma brucei*

1.5.1 General information on the disease

The African Sleeping Sickness is a disease caused by the parasite *Trypanosoma brucei*. Over the last 115 years, 3 major epidemics have spread over the African continent killing approximately 1 million people. Currently, it is estimated that around 30000 new infections arise per year, which are considered to be fatal if left without treatment [Hannaert, 2011]¹.

The life cycle of *Trypanosoma* involves the tsetse fly as a vector, which takes up the parasite together with human blood. In the fly *T. brucei* migrates from the intestines to the salivary glands, from where it can be transmitted again to a human that is bitten by the fly.

In the human host the sleeping sickness develops in 2 stages. During the first stage trypanosomes are located mainly in blood and lymph in order to promote transmission to its vector. Here it causes unspecific symptoms such as fever and headache. In the second stage the parasite crosses the blood-brain barrier and infects the central nervous system and other organs. This causes severe neurological effects, e.g. the name giving sleeping disorder.

1.5.2 Available treatments

The sleeping sickness has been subject to medical research from the beginning of the 19th century. Paul Ehrlich discovered the first trypanocidal substance in 1904, which led to the development of the first successful drug Suramin. Over the last century, new drugs have been introduced (see Table 1.2), but due to application restrictions, severe side effects, and resistance development Suramin is still in use. The most severe problem of the available drugs is that they all have to be administered by injection. Furthermore, some of the drugs, e.g. the combination of Eflornithine & Nifurtimox, have to be applied

¹<http://www.who.int/mediacentre/factsheets/fs259/en/>

1.5. GLYCOLYSIS IN *TRYPANOSOMA BRUCEI*

Table 1.1: List of currently applied trypanocidal drugs.

| Name | Year introduced | Treatment stage | Side effects | Resistance observed | Mode-of-action |
|-----------------|-----------------|-----------------|--------------|---------------------|----------------------|
| Suramin | 1916 | first | strong | no | promiscuous |
| Pentamidine | 1937 | first | few | yes ^a | unknown ^b |
| Melarsoprol | 1949 | second | toxicity | yes ^a | promiscuous |
| Eflornithine | 1990 | second | medium | maybe ^c | known ^d |
| E. + Nifurtimox | 1967 | second | few | maybe ^c | unknown ^e |

^a [Gehrig and Efferth, 2008]

^b binds mitochondrial deoxyribonucleic acid (DNA) [Barrett et al., 2007]

^c [Vincent et al., 2010]

^d irreversible inhibitor of ornithine decarboxylase [Bacchi and Yarlett, 1993]

^e contributes to production of reactive oxygen species [Enanga et al., 2003]

in a strict schedule, which makes medical personal necessary. This personal is not available in rural regions or in regions involved in warlike events. The Democratic Republic of the Congo is a good example for this. 70% of cases from the last decade were reported here and this can be seen as a result of the Second Congo War and its aftermath.

This and more detailed information can be found in recent reviews on trypanocidal drugs [Barrett et al., 2007, Hannaert, 2011].

1.5.3 Potential new treatments

As already mentioned, the available drugs for the treatment of the sleeping sickness are far from optimal. Since the geographic distribution of the disease is limited to the African continent by the distribution of the tsetse fly, the financial market for this drug is comparatively small. Thus, few commercial treatments have evolved and research on this topic is mainly driven by publicly funded institutions, which investigate how new trypanocidal drugs can be developed. As there is already a vast amount of data publicly available, including the pathogen's genome [Berriman et al., 2005], the development of potential drug candidates through public research seems to be feasible.

Unfortunately, the development of a vaccine against *T. brucei* appears to be unlikely. The pathogen is covered in so-called variable surface proteins, a set of approximately 1000 proteins, that is stochastically activated by DNA recombination [Morrison et al., 2009]. Therefore, the pathogen can easily avoid the human immune response and since prevention measures are not sufficient, a potential cure has to focus on treatment of the patient after infection.

Different pathways of trypanosomes have already been subject to research: like other rapidly dividing cells pathogens heavily rely on the cel-

1.5. GLYCOLYSIS IN *TRYPANOSOMA BRUCEI*

lular machinery for cell proliferation, which includes for example energy metabolism, biomass production, or cell cycle control [Hannaert, 2011].

Table 1.2: List of drugs and their targets in the extended trypanosomal glycolysis. Apart from the targets mentioned above, glyceraldehyde-3-phosphate dehydrogenase, aldolase [Perie et al., 1993], and phosphoglycerate kinase [Bernstein et al., 1998] have been inhibited *in vitro*.

| Enzyme | Drug | Reference |
|-------------------------------------------|------------------------|-----------------------------|
| Pyruvate kinase | suramin | [Morgan et al., 2011] |
| | melarsoprol | [Flynn and Bowman, 1974] |
| Hexokinase | suramin | [Wills and Wormall, 1950] |
| Glycerol-3-phosphate dehydrogenase (NAD+) | suramin | [Fairlamb and Bowman, 1977] |
| | cymelarsan | [Denise et al., 1999] |
| Phosphofructokinase | polycarpol | [Ngantchou et al., 2009] |
| Trypanosome alternative oxidase | salicylhydroxamic acid | [Clarkson et al., 1981] |
| | azaantraquinone | [Nok, 2002] |
| | ascofuranone | [Minagawa et al., 1997] |

One of these pathways, the glycolysis, seems to be particularly interesting for several reasons. First, while *T. brucei* is residing in the human bloodstream, glucose is its primary energy source and targets in glycolysis have proven to be essential in RNA (ribonucleic acid) interference experiments [Albert et al., 2005, Cáceres et al., 2010]. Second, many of the glycolytic enzymes have special structural and kinetics features distinguishing them from the human homologues, which simplifies the process of designing selective inhibitors against them [Verlinde et al., 2001]. Third, the spatial organisation of the glycolysis as seen in the pathogen is rarely seen in other organisms. The glycosome, a compartment similar to a peroxisome, accommodates the upper part and half of the lower part of the glycolysis. It has been proposed that this compartmentalisation is an alternative to feedback regulation of the pathway [Haanstra et al., 2008] and allows for high concentrations of the metabolites in glycolysis [Bakker et al., 1995, 1997]. Finally, drugs with targets in glycolysis have proven to be efficacious in experiments (see Table 1.2 for a list of drugs and their targets).

1.5.4 Available models of trypanosomal glycolysis

Glycolysis in *T. brucei* has already been subject to extensive investigation and modelling. The first model describing this pathway with ODEs was published in 1997 and was completely based on the literature and experimental data from single enzyme measurements [Bakker et al., 1997]. As this information was incomplete, some reactions were assumed to be in equilib-

1.5. GLYCOLYSIS IN *TRYPANOSOMA BRUCEI*

rium and several reactions were lumped together. The model was extended in 1999 including, for example, a dihydroxyacetone (DHAP) – glycerol-3-phosphate (G3P) – antiporter between glycosome and cytosol [Bakker et al., 1999]. Helfert *et al.* extended the model further by detailed kinetics for the triosephosphateisomerase (TPI) and the phosphoglucoisomerase (PGI) [Helfert et al., 2001]. In 2005 many of the model’s kinetics were updated using recent measurements of kinetic parameters and the last lumped reaction in the lower glycolysis was replaced by its explicit counterparts [Albert et al., 2005]. This model version provided the basis for work of Haanstra *et al.*, who created an alternative model version to study the effect of the compartmentalisation through the glycosome [Haanstra et al., 2008]. The most recent version of the trypanosomal glycolysis comprises fewer reactions than the 2005 version but it is accompanied by a website ² including different measurements of the model’s kinetic parameters. This, in principle, allows for investigations of the effects of parameter variations on the behaviour of the trypanosomal glycolysis [Achcar et al., 2012].

1.5.5 *In silico* determination of drug targets

The question for the most relevant drug targets in the trypanosomal glycolysis has so far been answered using different methods. The enzymes were compared on the basis of control of the catalysed reaction on the flux through the network (e.g. [Achcar et al., 2012]) or based on dynamic simulations of the flux change in response to a reduction in a reaction’s maximal velocity (e.g. [Bakker et al., 1999]). Latter results were compared to experiments in which the maximal velocities were modulated by reducing the enzymes’ concentrations through RNA interference [Fire et al., 1998].

The second and the third version of the model supported the idea that the most potent targets in the glycolysis are the trypanosomal hexose transporter (THT), fructose bisphosphate aldolase (ALD), glycerol-3-phosphate dehydrogenase (GPDH), glyceraldehyde phosphate dehydrogenase (GAPDH), and phosphoglycerate kinase (PGK) [Bakker et al., 1999, Helfert et al., 2001]. Later model versions support the idea that THT is the most potent target followed by GAPDH and phosphoglycerate mutase (PGM). Furthermore, PGK has been predicted to be less promising while the enolase (ENO) has received more attention [Albert et al., 2005, Haanstra et al., 2008, Achcar et al., 2012].

²<http://silicotryp.ibls.gla.ac.uk/wiki/Glycolysis>

1.6 The arachidonic acid pathway

1.6.1 Physiology

1.6.1.1 Pathway structure

A second example system that I will investigate throughout this thesis is the arachidonic acid (AA) pathway in humans. The AA pathway, as it is depicted in Figure 1.1, describes the conversion of AA into various eicosanoids. It starts with the release of AA from membrane phospholipids and diacylglycerols by the phospholipases A₂ (PLA₂), C, and D (mainly cytosolic phospholipase A₂ α [Ghosh et al., 2006]) and diacylglycerol lipases [Farooqui et al., 1997]. AA is then converted via different pathways to prostaglandins (PGs), prostacyclin (PGI₂), thromboxanes (TXs), and leukotrienes (LTs). These fatty acids act as autocrine and paracrine signalling molecules, binding to G-protein coupled and nuclear receptors, activating various signalling cascades leading to diverse responses.

Under normal conditions, eicosanoids are involved in inflammation, vascular homeostasis, and the protection of the gastric mucosa [Harizi et al., 2008]. While PGE₂ is synthesised by most human cells, other eicosanoids are mainly synthesised in specific cells or tissues (e.g. thromboxanes in platelets and macrophages or leukotrienes in leukocytes, macrophages, and mast cells) [Funk, 2001]. Thromboxane A₂ (TXA₂) promotes platelet coagulation and acts as a vasoconstrictor, which is counteracted by the vasodilator PGI₂ [Marcus, 1978]. LTB₄ plays a role in the immune response by acting as a chemoattractant [Weller et al., 2005]. Together with PGE₂, which induces fever and acts in a proinflammatory way [Ivanov et al., 2004], it is the most relevant metabolite of the arachidonic acid pathway. Further eicosanoids, e.g. epoxyeicosatrienoic acids (EETs), play a role in inflammation and cell proliferation [Zeldin, 2001], but they are supposed to be of lesser importance and are not described in detail in this context. For a more complete overview on the list of actions of eicosanoids the reader is referred to [Harizi et al., 2008].

1.6.1.2 Regulation

The production of inflammatory mediators via the AA pathway is a heavily controlled process and this control is exerted via the activation of proteins as well as changes in transcription. cPLA₂ α is catalysing the first reaction of the pathway and is supposed to exert most control over the production of eicosanoids [Wymann and Schneider, 2008]. The enzyme is activated in response to a proinflammatory stimulus, which induces a MAP kinase sig-

1.6. THE ARACHIDONIC ACID PATHWAY

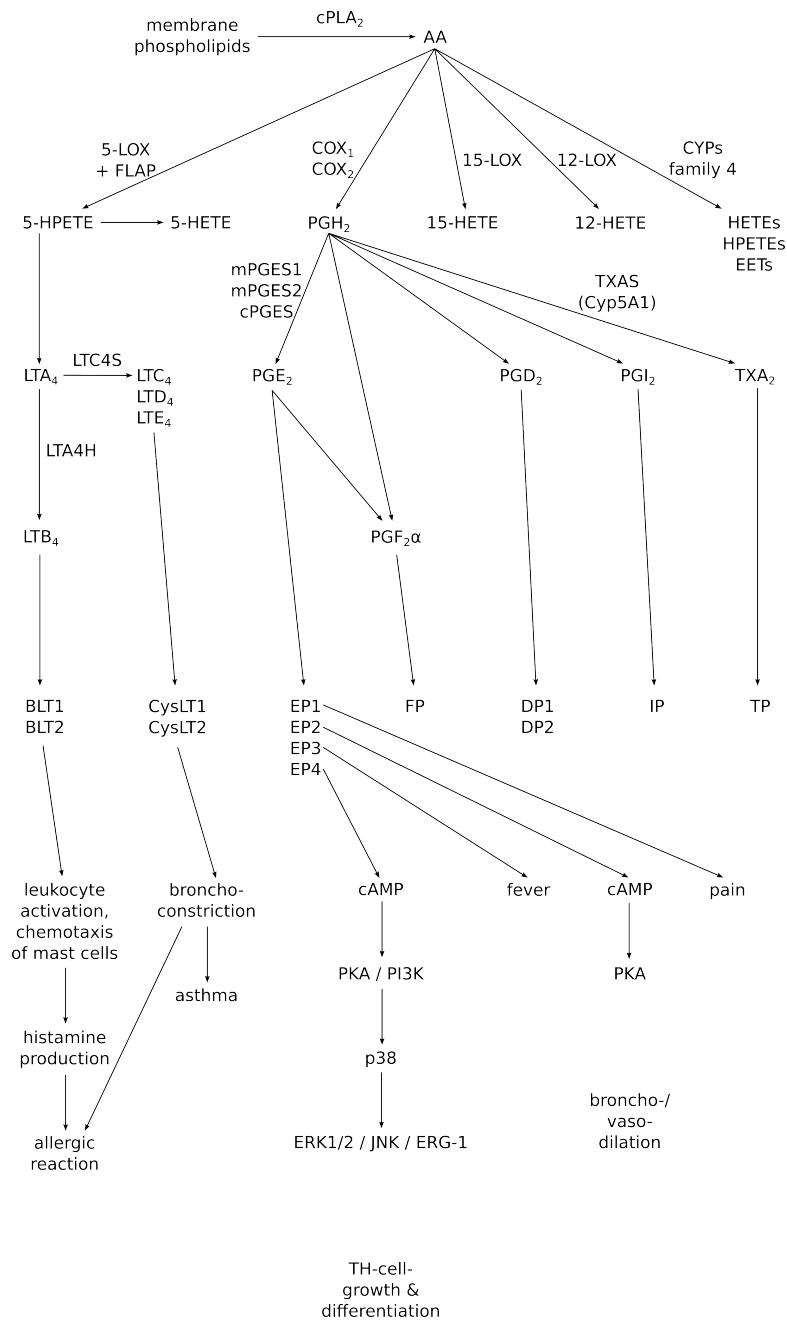


Figure 1.1: The arachidonic acid pathway, its regulation, and its downstream effects. Arcs in the graph either describe chemical conversion, receptor binding, or promotion of a physiological effect. This Figure has been compiled from various resources mentioned in this section, foremost [Yang et al., 2008].

1.6. THE ARACHIDONIC ACID PATHWAY

nalling cascade and the release of calcium. Calcium binds first to cPLA₂ α and induces its translocation into membranes (like ceramide-1-phosphate [Pet-tus et al., 2004]). Afterwards it is phosphorylated by ERK1, ERK2, p38, calmodulin protein kinase II, and Mnk1, which further enhances its activity [Bonventre, 2004]. Furthermore, the transcription of cPLA₂ α is induced by cytokines such as Interleukin-1 α or tumour necrosis factor [Clark et al., 1995], e.g. via the JAK-STAT-pathway [Neeli et al., 2004], or through JNKs and ERKs [Van Putten et al., 2001].

The arachidonic acid pathway is not only controlled on the level of cPLA₂ α , however, many of the enzymes are regulated in a similar way. The microsomal PGE₂ synthase-1 (mPGES1) is for example induced by the same cytokines [Stichtenoth et al., 2001] and the transcriptional regulation of 5-lipoxygenase (5-LOX) [Rådmark et al., 2007] and cyclooxygenase 2 (COX-2) [Reddy et al., 2000] is similar to the regulation of cPLA₂ α . This coregulation of the first enzymes in the arachidonic acid pathway might indicate that cPLA₂ α is not the only rate limiting enzyme, but that other enzyme concentrations need to be raised as well to induce eicosanoid production [Herschman et al., 1997].

1.6.1.3 Downstream effects

After eicosanoids have been produced via the AA pathway, they leave the cells through so-called multidrug resistance-associated proteins and act in an autocrine or paracrine fashion on G-protein coupled receptors. The receptors and their preferred ligands are shown in Figure 1.1. IP, DP, EP₂, and EP₄ are so-called “relaxant” receptors, exerting their action through increasing cAMP, EP₁, FP, and TP are “contractile” receptors, signalling through releasing calcium, and EP₃ is an “inhibitory” receptor mediating fever through decreasing cAMP levels [Ushikubi et al., 1998, Funk, 2001]. Downstream effects of the activated receptors include pain mediation through EP₁ [Stock et al., 2001], induction of differentiation and growth in certain T helper cells through EP₄ [Yao et al., 2009], and activation as well as attraction of other immune system cells through BLTs [Tager and Luster, 2003, Lundeen et al., 2006]. Downstream of the activated receptors a number of different signalling pathways are activated, e.g. Wnt, EGFR-PI3K-AKT, and MAP kinase pathways [McCarty, 2004, Cha and DuBois, 2007]. For a recent overview on those pathways the reader is referred to [Wang and DuBois, 2010].

Except from actions on GPCRs, eicosanoids have been shown to bind to peroxisomal proliferator-activated receptors (PPARs), which are located in the nucleus, although the involvement of PPARs in the *in vivo* action of eicosanoids is controversially discussed [Funk, 2001].

1.6. THE ARACHIDONIC ACID PATHWAY

1.6.2 Pathophysiology

1.6.2.1 Diseases involving the AA pathway

Inflammation As already mentioned, the arachidonic acid pathway is the central metabolic part of the mediation of inflammatory processes. Therefore, its deregulation can lead to uncontrolled inflammatory responses which are implicated in autoimmune diseases and allergies but are also supposed to play a role in diseases such as atherosclerosis, asthma, Alzheimer's, Parkinson's, and certain types of cancer [Serhan et al., 2008]. The inflammatory response is mainly supposed to be caused by PGE_2 , which is the most abundant prostaglandin and elevated in almost all of the aforementioned diseases [Ushikubi et al., 1998, Samuelsson et al., 2007].

Asthma The symptoms of asthma are supposed to be mainly caused by cysteinyl leukotrienes via the CysLTs, which induce bronchoconstriction, but also the balance of the eicosanoids $\text{PGD}_2 + \text{TXA}_2 / \text{PGE}_2 + \text{PGI}_2$ seems to play a role [Wenzel, 1997]. The increase in the leukotrienes is caused by a change in the composition of the T-helper cell population and the fact that different types of T-helper cells produce eicosanoids to a different extent [Robinson et al., 1992]. Furthermore, the leukotriene LTB_4 plays an important role in asthma via attraction of immune response cells and acting pro-inflammatory [Peters-Golden and Henderson Jr, 2007].

Cancer As general promoters of inflammation prostaglandins and leukotrienes are furthermore assumed to play a role in the development of cancer [Wang and DuBois, 2010]. Tumor growth is assumed to be accompanied by a change in its microenvironment, where an inflammation attracts leukocytes, which in turn produce more eicosanoids. If this positive feedback is not inhibited, as it is in healthy tissue, chronic inflammations can occur [Mantovani et al., 2008, Serhan et al., 2008].

1.6.2.2 Drugs acting in the AA pathway

Because of the central role of the AA pathway in many relevant diseases, there is a high commercial interest in drugs acting on this pathway. Not without reason acetylsalicylic acid has been described as “one of the most durably successful commercial products of all time” [Jeffreys, 2010]. After its introduction to the market in 1899 under the brand name “Aspirin” it has been used in the treatment of pain, fever, and various inflammatory diseases. Furthermore, it is used as a treatment after infarction [Antithrombotic Trialists' Collaboration, 2002] and its long-term use has been shown to reduce

1.6. THE ARACHIDONIC ACID PATHWAY

death rates due to cancer [Rothwell et al., 2011]. Unfortunately, its use has been associated with an increased risk of gastrointestinal bleeding, which partially reduces its application areas [Derry and Loke, 2000]. On the molecular level aspirin exerts its effect by irreversibly binding and inhibiting COX-1, the “house-keeping” isoform of the prostaglandin G_2 / H_2 synthase [Roth and Majerus, 1975]. This effect is most dominant in platelets, which cannot produce new enzymes and, therefore, become unable to produce TXA_2 , causing the side-effects.

In order to circumvent the side-effects of aspirin or other NSAIDs (non-steroidal anti-inflammatory drugs) which affect COX-1, drugs selectively targeting the inducible COX-2 isoform have been developed [Chan et al., 1999]. Unfortunately, COX-2 is the main COX through which PGI_2 is produced [FitzGerald, 2003], and reduced PGI_2 levels lead to high blood pressure, atherosclerosis, and thrombosis [Wong et al., 2005, Yu et al., 2012]. Accordingly, selective COX-2 inhibitors, such as rofecoxib (Vioxx), are associated with an increased heart attack risk [FitzGerald, 2004, Jüni et al., 2004, Furberg et al., 2005] and have thus been taken off the market again. A second potentially relevant mechanism by which the side-effects of selective COX-2 inhibitors can be explained come from a secondary role of the cyclooxygenases. In later phases of the response to a stimulus, COX-2 is involved in the production of inflammation resolving mediators such as lipoxins, resolvins, and protectins, whose production is delayed under COX-2 inhibition [Gilroy et al., 1999].

Apart from the cyclooxygenases, other enzymes and receptors have been targeted in the search for potent anti-inflammatory drugs. A short summary on some substances with an inhibitory effect on the AA pathway and their targets is shown in Table 1.3. This table shows that the AA pathway is offering a large number of targets that can be inhibited by small chemical entities.

1.6.3 Modelling the AA pathway

In order to understand the effects and side-effects of available drugs and to contribute to the development of new drugs, the arachidonic acid pathway has been subject to modelling efforts. The first dynamic model of AA metabolism has been developed by Yang *et al.* [Yang et al., 2007]. This model described the pathway as present in human polymorphonuclear leukocytes and comprised the production of LTB_4 , PGE_2 , thromboxanes and 5, 12, and 15-HETE (hydroxyeicosatetraenoic acid). Afterwards, the model has been extended to include the production of PGI_2 and to incorporate the metabolism in platelets and endothelial cells [Yang et al., 2008]. With

1.7. OUTLINE OF THIS WORK

Table 1.3: Potential targets in the AA pathway, their inhibitors or antagonists, and the diseases which are supposed to be treated with them. Abbreviated terms include diacylglycerol lipase (DAGL), 5-LOX activating protein (FLAP), arachidonyl trifluoromethyl ketone (ATK), cardiovascular diseases (CV), and Alzheimer’s disease (AD).

| Target | Drug | Treated disease | Reference |
|-----------------------------------------|----------------------|---------------------------------------|----------------------------------------|
| cPLA ₂ α ^a | dexamethasone | inflammation | [Clark et al., 1995] |
| | giripladib | arthritis | [Suckling, 2010] |
| COX-1 | acetylsalicylic acid | pain, inflammation, fever, thrombosis | |
| COX-2 ^b | valdecoxib | arthritis | [Hood et al., 2003] |
| 5-LOX | zileuton | asthma | [Carter et al., 1991] |
| FLAP ^d | licofelone | arthritis | [Koeberle et al., 2008] |
| mPGES-1 | MF63 | pain & fever | [Côté et al., 2007a] |
| LTA4H | bestatin | cancer | [Peters-Golden and Henderson Jr, 2007] |
| | 6-gingerol | cancer | [Jeong et al., 2009] |
| 12-LOX | baicalein | cancer & AD | [Sekiya and Okuda, 1982] |
| EP1 | ONO-8711 | cancer | [Kawamori et al., 2001] |
| EP2 | AH 6809 | cancer | [Woodward et al., 1995] |
| EP4 | AH23848 | inflammation | |
| | ONO-AE3-208 | cancer | [Fulton et al., 2006] |
| BLT1/2 | LY293111 | asthma | [Evans et al., 1996] |
| CysLT1 | zafirlukast | asthma | [Hui et al., 2001] |
| TP&DP2 | ramatroban | atherosclerosis | [Terada et al., 1998] |
| FP | bimatoprost | ocular hypertension | [Woodward et al., 2001] |

^a expression is reduced

^b selectivity is achieved through slow dissociation (pseudo-irreversibility)

^c also a mPGES-1 inhibitor

^d also targets COXs

this extension the model was able to reproduce the effects and side-effects of various NSAIDs *in silico*.

In total, the latest model includes 117 parameters (excluding substrate and enzyme concentrations) from which 46 have been determined in experiments. The rest of the parameters has been fitted using sparse experimental data, which left many of the parameters unconstrained. In order to deal with the resulting uncertainty about the outcome of simulations of this model, the authors have accompanied the model with several parameter sets describing the experimental data equally well. However, it has been concluded that parameter variations only had a minor effect on the qualitative outcome of inhibitors in the network [Yang et al., 2008].

1.7 Outline of this work

1.7.1 The big picture

Over the last years the rate of drug discovery has declined significantly, which can, in the case of drugs developed following a target-based approach, be

1.7. OUTLINE OF THIS WORK

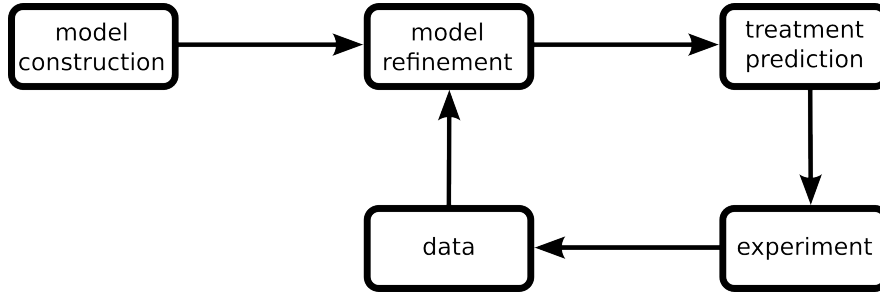


Figure 1.2: Overview on the workflow of TIDE for the identification of potent targets with the help of mathematical models.

attributed to the drugs' lack of efficacy or their toxicity in clinical trials. In order to tackle this problem I apply methods from Systems Biology to the identification of potent drug targets, which will help in the development of effective and safe drugs.

For this purpose I have developed a framework, as visualised in Figure 1.2, which applies the cycle of Systems Biology to the drug target identification problem. First, a mathematical model is constructed and refined using available experimental data. Then, predictions on drug actions are made from the model and tested in experiments, that support or invalidate the hypotheses. And finally, data that does not agree with the model can be used to refine it again such that new predictions can be made and the cycle starts anew.

In contrast to other modelling efforts, the predictions in my framework mainly focus on the actions of drugs on the network. While the idea of modelling drug actions is not new in general, published methods do follow slightly different approaches implementing it. My work aims at the unification, formalisation, and extension of these methods as well as at the development of tools supporting scientists to follow my framework.

1.7.2 Contents of this thesis

This thesis, which introduces the aforementioned methods and tools, is divided into different chapters. Each of them describes the methods than can be used in different stages of the framework. I start with a general introduction of ODE models and mathematical methods for their storage, analysis, and refinement in chapter 2. Then, I introduce methods for the retrieval, analysis, and refinement of models that can be used in the construction of a mathematical model, which is later used in the target identification step. Chapter 3 introduces similarity measures for models and data sets, which aid

1.7. OUTLINE OF THIS WORK

in the search for relevant available knowledge, while chapter 4 extends those measures to help in the alignment and annotation of models. These methods can be used in the construction of large comprehensive dynamic networks. In the following part of the thesis, I introduce different ideas of how mathematical models can be used for the prediction of targets. For this purpose I formalise the target identification as a parameter estimation problem in chapter 5 and apply the concepts of network selectivity and non-identifiability to it. Furthermore, in chapter 6 I search for synergisms and antagonisms across different drug targets. Finally, I discuss the impact of my work and its possible extensions in chapter 7. Throughout this thesis, parts of the framework are applied to small example models as well as to two biologically relevant diseases, the African sleeping sickness and the inflammatory response. The results of these analyses highlight new, potentially interesting experiments, which can be used to improve the predictive power of the used mathematical models.

Chapter 2

Methods

Contents

| | | |
|------------|--------------------------------------------------------|-----------|
| 2.1 | Modelling using ordinary differential equations | 47 |
| 2.1.1 | Problem formulation | 48 |
| 2.1.2 | Solutions in time | 48 |
| 2.1.3 | Structure of models | 49 |
| 2.1.4 | Metabolic control analysis | 51 |
| 2.2 | Parameter estimation | 52 |
| 2.2.1 | Problem formulation | 52 |
| 2.2.2 | Algorithms and heuristics for optimisation | 53 |
| 2.2.3 | Statistical assessment of fit quality | 54 |
| 2.3 | Model representation | 54 |
| 2.3.1 | Systems Biology Markup Language | 54 |
| 2.3.2 | Model annotation | 55 |
| 2.3.3 | Model repositories | 56 |

2.1 Modelling using ordinary differential equations

Because this work is mainly concerned with the application of mathematical models of biochemical processes to the problem of the identification of efficient and safe drug targets, I will introduce basic concepts necessary for the development of such models within this chapter. This introduction is

2.1. MODELLING USING ORDINARY DIFFERENTIAL EQUATIONS

based on [Schulz and Klipp, 2010], and it will cover the mathematical formulation of ordinary differential equation (ODE) models, their construction, their analysis, and their representation.

2.1.1 Problem formulation

ODE models in general, describe the time evolution of state variables by differential equations. This means that for a vector of state variables $y(t)$ an ODE model describing how the variables change over time is given by a vector of initial concentrations $y(t_0)$ and a vector of differential equations $\frac{d}{dt}y(t) = f(y(t), \theta, t)$. The differential equations are allowed to refer to the current values of the states variables $y(t)$ as well as to a vector of model parameters θ and the current time t .

In the context of Systems Biology the state variables usually describe concentrations or molecule numbers of various substances. These substances can be small metabolites as well as macromolecules like proteins, RNA, or DNA. The differential equations describe the processes converting substances and the parameters determine the velocities of these conversions. In biological contexts, these equations are usually not explicitly dependent on the time t . This is a special case of ODEs that is termed autonomous. It should be noted that ODE models are only applicable to systems in which one deals with large molecule numbers and which are supposed to be well stirred. If these assumptions are not true, one should either use a stochastic simulations or incorporate spatial aspects of the system.

2.1.2 Solutions in time

The problem of determining the concentrations of substances at a certain time $y(t)$, given starting values $y(t_0)$ and the differential equations f , has been called the initial value problem (IVP). For large systems this problem can rarely be solved analytically. Exceptions to this are for example linear models, e.g. $\frac{d}{dt}y = a \cdot y(t)$ which has the general solution $y(t) = y(t_0) \cdot e^{at}$. In most cases, however, the dynamic behaviour of the system has to be approximated analytically.

The most simple numerical approximation of an IVP can be performed using the Euler method [Euler, 1768]. For a given step size h the method works by successively calculating the state $y(t+h)$ from $y(t)$, resulting in a set of time points $y(t_0 + i \cdot h)$. The main assumption of the Euler method is that the differential equations f do not significantly change within a time window $[t, t+h]$ and that they can therefore be approximated by $f(y(t), \theta, t)$. As a result, the time evolution of the system can be approximated by the iterative

2.1. MODELLING USING ORDINARY DIFFERENTIAL EQUATIONS

calculations

$$y(t + h) = y(t) + h \cdot f(y(t), \theta, t). \quad (2.1)$$

Euler's method, however, is rarely used anymore, because its numerical accuracy only decreases linearly with the step size and because the method has severe problems with the stability of its solutions. Stability in the context of ODE solvers refers to the fact under which conditions a numerical solution converges to a stable steady state. These can for various ODE solvers be different to the conditions under which the analytical solution converges. Nevertheless, more complex methods are built upon this simple numerical scheme and iteratively calculate new time points from old ones. Notable improvements to this general idea are implicit methods, which allow $y(t + h)$ to appear on the right hand side of Eq. 2.1, Runge-Kutta methods [Runge, 1895], evaluating f at time points outside the grid $t = t_0 + i \cdot h$, and Adams-Bashford methods [Bashforth and Adams, 1883], considering multiple time points calculated in previous steps.

2.1.3 Structure of models

2.1.3.1 Stoichiometric matrix

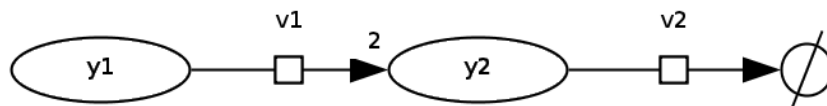


Figure 2.1: Graphical representation of the stoichiometry of the example network.

When dealing with ODE models describing large networks it becomes useful to divide the differential equations into two parts, the network's stoichiometry N and a vector of reaction velocities $v(y(t), p)$. The differential equations then read $\frac{d}{dt}y(t) = N \cdot v(y(t), p)$. N , the stoichiometric matrix, has the dimension $|y| \times |v|$, where $|y|$ is the number of variables of the system and $|v|$ is the number of reactions. Its entries describe whether an entity is produced or consumed by a reaction via a positive or a negative sign and how many molecules are converted per elementary reaction step via their absolute values. For the simple example system shown in Figure 2.1 for example the differential equations read

$$\frac{d}{dt}y(t) = \begin{pmatrix} -1 & 0 \\ 2 & -1 \end{pmatrix} \cdot v(y(t), p).$$

2.1. MODELLING USING ORDINARY DIFFERENTIAL EQUATIONS

This notation is advantageous for theoretical considerations as well as numerical calculations, as complex kinetic expressions only need to be evaluated once instead of for every variable being changed by it.

2.1.3.2 Kinetics

Reaction velocities, the functions $v(y(t), p)$, describe how fast processes happen in the considered system. In this context, reactions do not only describe reactions in the classical chemical sense, but all kinds of processes converting or transporting molecules in a system. Depending on the type of process described by a reaction and the amount of knowledge one has about factors influencing its velocity, different kinds of formulas are used.

Constant rates are the most simple reaction kinetics and are used for processes creating matter that are not understood well enough to be described with a more complex formula.

Mass action kinetics regard the reaction velocity to be proportional to the concentrations of the substrates to the power of their stoichiometry [Guldberg and Waage, 1864]. They follow the general formula

$$v_j = a \cdot \prod_{i|N_{ij}<0} y_i(t)^{-N_{ij}} - b \cdot \prod_{i|N_{ij}>0} y_i(t)^{N_{ij}}.$$

Michaelis-Menten type kinetics are compact descriptions of enzymatic reactions [Menten and Michaelis, 1913]. Their formulas are derived from detailed models of elementary reaction steps, which describe the binding and dissociation of substrates and products as well as the enzymatic catalysis, and are based on the following assumptions. First, for an irreversible reaction product formation and release is irreversible and slow compared to the formation of the substrate-enzyme-complex. Second, the enzyme concentration is low compared to the concentration of substrates and products. Third, the concentration of the enzyme-substrate complex is supposed to be in steady state. The final formula for the rate of product formation in an irreversible enzymatic reaction j with one substrate i is given by

$$v_j = \frac{V_{\max} \cdot y_i(t)}{K_m + y_i(t)},$$

where V_{\max} is the maximal reaction velocity and K_m (the Michaelis-Menten constant) is the substrate concentration at which the reaction assumes half its maximal speed. More details on these kinetics and a visualisation of the reaction scheme are given in section B.2.1.1 in the Appendix.

2.1. MODELLING USING ORDINARY DIFFERENTIAL EQUATIONS

Other kinetics than the aforementioned are known in the literature. They have been deduced using similar approaches and extend to cases in which reactions are reversible, the reactions involve multiple substrates and products that bind in a certain order, and the substrate binding is cooperative, i.e. the association of one molecule makes others more likely to bind. Furthermore, kinetics can be extended to respect inhibitors or activators of enzymatic reactions. For Michaelis-Menten kinetics one can distinguish different inhibition types as shown in section B.2.1 in the Appendix. These different types depend on the conformation of the enzyme the inhibitor is able to bind. A competitive inhibitor can only bind the free enzyme, an uncompetitive one only the substrate-enzyme-complex, and a noncompetitive inhibitor binds the enzyme regardless of its conformation.

2.1.4 Metabolic control analysis

2.1.4.1 Steady state

For some applications it is not necessary to study the dynamic behaviour of a system in full detail. Instead one can investigate the approximate behaviour for long timescales. For times approaching infinity the system can generally behave in three different ways. It can converge to a single point in state space, which is the space spanned by the state variables, it can converge to a cyclic trajectory, or it can not converge at all.

In the first case the point in state space towards which the system is driven is called an attractor or a stable steady state. For steady states $y^*(t)$ the equation $\frac{d}{dt}y^*(t) = 0$ holds, which means that once the system has assumed this state, it is unable to leave it on its own. If a model has stable steady states, trajectories starting in the close proximity always converge to this attractor over time.

2.1.4.2 Elasticities

Metabolic control analysis (MCA) has been developed to study changes in models' steady state behaviour in response to changes in parameters or initial concentrations [Kacser and Burns, 1973, Heinrich and Rapoport, 1974]. In MCA global properties of a system, the control and response coefficients, are computed from the stoichiometry of the network and local properties of the reactions, the elasticities. The normalised ε - and π elasticities describe the change in reaction velocity in response to changes in substrate concentrations

2.2. PARAMETER ESTIMATION

or parameter values as defined by

$$\varepsilon_i^j = \frac{\partial \ln v_j}{\partial \ln y_i} = \frac{y_i}{v_i} \frac{\partial v_j}{\partial y_i} \quad (2.2)$$

$$\pi_m^j = \frac{\partial \ln v_j}{\partial \ln p_m} = \frac{p_m}{v_i} \frac{\partial v_j}{\partial p_m}. \quad (2.3)$$

2.1.4.3 Control and response coefficients

Through the connection of N , ε , and π , global changes in the steady state of a reaction system can be computed. In this steady state the concentrations are given by the vector y^* while the reaction velocities are termed fluxes and abbreviated with J . The influence of reaction velocities and parameter values on the steady state concentrations and fluxes is given by the normalised flux and concentration control coefficients and the normalised flux and concentration response coefficients

$$C_j^k = \frac{v}{J} \frac{\partial J / \partial p}{\partial v / \partial p} \quad (2.4)$$

$$C_j^i = \frac{v}{y^*} \frac{\partial y^* / \partial p}{\partial v / \partial p} \quad (2.5)$$

$$R_m^k = \frac{p}{J} \frac{\partial J}{\partial p} \quad (2.6)$$

$$R_m^i = \frac{p}{y^*} \frac{\partial y^*}{\partial p}. \quad (2.7)$$

For details on how the coefficients can be deduced from the stoichiometric matrix and the elasticities the reader is referred to [Klipp et al., 2009, Ch. 2].

2.2 Parameter estimation

The full description of a kinetic model includes its stoichiometry, its kinetics, and numerical values for parameters and starting concentrations. Some parameters and concentrations can be obtained from public web resources as BRENDA [Scheer et al., 2011], but the remaining ones need to be guessed. This guessing is performed in such a way that the dynamical model is able to reproduce experimental data and is called parameter estimation.

2.2.1 Problem formulation

In the parameter estimation process a few assumptions are made. First of all, one assumes that a deterministic process is underlying the experimentally

2.2. PARAMETER ESTIMATION

obtained data. Second, this process can be described by a kinetic model with a certain parameter set. And third, the experimental data is in principle determined by the model but additionally contains some independent normally distributed noise. Under these assumptions the likelihood of the experimental data given the correct model can be described by a normal distribution. This likelihood can be shown to have the same extremal points with respect to parameter values as the likelihood of the model given the data. Furthermore, the likelihood can be linked to the so-called objective function, which is given by

$$\mathcal{X}^2 = \sum_{i,j} \left(\frac{\bar{y}_i(t_j) - y_i(t_j, \theta)}{\sigma_i(t_j)} \right)^2,$$

where \bar{y} describes experimentally measured data points and σ is the standard deviation of the measurements, and it can be shown that a parameter set minimising \mathcal{X}^2 is a maximum likelihood estimate of the parameters given the data. Mathematical details of this reasoning are given in section B.2.3.1 in the Appendix.

2.2.2 Algorithms and heuristics for optimisation

While one now knows what function should be minimised in order to get a good estimate of parameters in agreement with experimental data, it can be discussed how this objective function can be minimised. The methods with which such an optimisation can be performed fall into two different categories.

On the one hand there are local methods, like gradient-descent or BFGS [Broyden, 1970, Fletcher, 1970, Goldfarb, 1970, Shanno et al., 1970]. The methods are algorithms that find a point in parameter space with the lowest objective function value in close proximity to an initial starting point. If one imagines the objective function as a surface, they try to find the lowest point in the valley that contains the starting position. These methods are reliable and fast in identifying local optima, however, their results depend on the initial starting point and cannot be assumed to be optimal along the complete parameter space.

On the other hand there are global methods which aim at finding optimal points in the complete parameter space. Because the objective function cannot be assumed to have a particular shape, global methods need to evaluate lots of parameter sets from different regions of parameter space. For this purpose they remember one or a population of (sub-)optimal solutions from which they jump to points in parameter space that will be evaluated next in a particular way. Examples of such methods are simulated annealing

2.3. MODEL REPRESENTATION

[Kirkpatrick et al., 1983] and differential evolution [Storn and Price, 1997]. Depending on the particular problem that is investigated, different global optimisation heuristics might be successful in identifying the most favourable parameter set. However, these methods can also not guarantee to find the globally optimal solution within a limited time.

2.2.3 Statistical assessment of fit quality

Just given the value of the objective function one is not generally able to decide how good a model fits the experimental data in absolute terms. Nevertheless, the objective function can be used to judge which model from a series of models is the most probable one given the experimental data. Such an analysis can be performed using likelihood ratio tests, as long as the models are nested. For other kinds of compared models the Akaike information criterion [Akaike, 1974] can be applied. This criterion assigns each model a score based on the fit quality and its number of parameters and allows to judge whether a more complex model is really needed to explain the data.

2.3 Model representation

In order to facilitate the easy reuse of mathematical models, different standards for the representation of dynamic models have been developed. The implementation of a model in such a standardised format enables its analysis in a large number of computational tools and is therefore preferable. However, not all models can be saved in them as they are restricted in their expressivity. For a general guide on how reusable models are created the reader is referred to [Krause et al., 2011].

2.3.1 Systems Biology Markup Language

The Systems Biology Markup Language (SBML) [Hucka et al., 2003] is an XML-based exchange format for dynamical models of reaction networks. Internally it is organised as lists of different entities, e.g. compartments, species, which are the main variables, reactions, which convert species into each other, and parameters, which are numerical values used in formulas such as kinetic laws, assignments, or triggered events. By now, SBML is supported by more than 200 tools ¹ including tools for construction, visualisation, simulation, management, conversion, and various types of analyses. Furthermore, it

¹<http://sbml.org>

2.3. MODEL REPRESENTATION

supports various kinds of model formalisms. Stoichiometric models are supported as well as dynamic ones and the latter can be formulated in either a deterministic or in a stochastic way.

Other languages for the representation of models exist, however, they are less well suited to describe the kind of kinetic models investigated within this work. CellML [Lloyd et al., 2004] is a very general XML-based exchange format, which does in principle not only allow the description of biochemical processes but can represent any kind of mathematical model. However, CellML models do not necessarily represent a model's stoichiometry, which is required for my further analyses, in an unambiguous way. BioPAX (Biological Pathway Exchange) [Demir et al., 2010] is an RDF/OWL-based format for the representation of pathway related information. In contrast to SBML this format is lacking quantitative descriptions of the dynamic behaviour of the model, which is required by my methods to judge the quality of different targets.

2.3.2 Model annotation

Proper access to the differential equations stored in a model is provided through the use of a standard model format. Using such a format, however, does not imply that human readers will be able to assess the biological content of a model, i.e. the physical entities or processes behind the model's variables. To address this problem annotations can be used to assign semantic information to model elements, which unambiguously identifies their meaning by stating a relation between the element and an entry from a web resource providing a controlled vocabulary (CV).

In order to enhance the value provided by annotations the MIRIAM standard [Le Novère et al., 2005] has been developed. This standard describes how semantic information can be incorporated into a model, which information should be provided, what kind of web resources can be linked to, and which formal relations can be used to describe certain biological facts.

The elements of a model can be linked to entries from various CVs including Gene Ontology [Ashburner et al., 2000], ChEBI [Degtyarenko et al., 2008], or UniProt [Bairoch et al., 2009]. By relating elements to entries from these resources, it is possible to describe all sorts of physical entities within a model, but none of these CVs can be used to assign a mathematical meaning to a parameter or a kinetic law. This is helped by a special CV, the Systems Biology Ontology (SBO) [Le Novère, 2006], which enables a model creator to precisely define the functional meaning of a model element.

2.3. MODEL REPRESENTATION

2.3.3 Model repositories

Standards and annotations increase the reusability of a given model. A final factor in the promotion of kinetic models is the availability of different databases storing them. The most popular example of such a database is BioModels Database [Le Novère et al., 2006], which is the largest collection containing hundreds of models. Another example is JWS online [Olivier and Snoep, 2004], which is much smaller but provides the opportunity to simulate a model's behaviour online. For information on further model repositories the reader is referred to PathGuide [Bader et al., 2006], a meta web resource on databases containing pathway information.

Part I

Finding and refining available biological knowledge

Chapter 3

Semantic similarity measures for Systems Biology

Contents

| | | |
|------------|-------------------------------------------------------------------------------------|-----------|
| 3.1 | Semantic information in Systems Biology | 60 |
| 3.1.1 | Using available information for modelling | 60 |
| 3.1.2 | Integrating ontologies | 65 |
| 3.1.3 | Similarity measure for entries in ontologies | 67 |
| 3.1.4 | Principles of information retrieval | 70 |
| 3.1.5 | Standards for semantic information on the internet | 70 |
| 3.2 | Comparing models and data sets based on se- mantic information | 71 |
| 3.2.1 | Combining ontologies | 71 |
| 3.2.2 | Similarity measures for Biological Concepts | 75 |
| 3.2.3 | Similarity measures for annotated data sets and models | 79 |
| 3.3 | Retrieval, alignment, and clustering of models and data sets | 83 |
| 3.3.1 | Assessing the quality of different similarity measures | 83 |
| 3.3.2 | Evaluation of the different measures | 84 |
| 3.3.3 | Matching experimental data to models | 87 |
| 3.3.4 | Matching Systems Biology models | 89 |
| 3.4 | Discussion | 92 |
| 3.4.1 | Retrieval of models and data sets | 92 |

3.1. SEMANTIC INFORMATION IN SYSTEMS BIOLOGY

| | | |
|-------|---------------------------------------------|----|
| 3.4.2 | Criteria for model similarity | 94 |
| 3.4.3 | Quality of the presented measures | 95 |
| 3.4.4 | Limitations of the current method | 96 |
| 3.4.5 | Conclusion | 96 |

3.1 Semantic information in Systems Biology

3.1.1 Using available information for modelling

Systems Biology models in drug target identification In the search for good drug targets which will lead to drugs with a high efficacy and a low toxicity one requires Systems Biology models of high quality. This high quality implies that the model has been constructed in a bottom-up fashion. Here, the modelling starts from elementary compounds and reactions which are put together to form a large comprehensive model including detailed kinetic knowledge. Otherwise, if the model has been constructed in a top-down or middle-out fashion, it renders necessary that relevant parts of the model are described to the detail of single enzymatic or non-enzymatic reactions. Furthermore, it implies that the model has been validated and refined using various kinds of high quality biological data. While the first requirement ensures the applicability of simple methods to simulate inhibition or activation of single reaction steps in the network, the second requirement increases the plausibility of the model's predictions. Nevertheless, it can never be guaranteed that a prediction is correct. A model will in most cases only be applicable for the purpose it has been designed for. Therefore, a model used for target prediction should have proven its predictivity by reproducing preferably many, diverse data sets, that are relevant to the treated disease.

But, even when the predictions made by a model are not fully correct, this model might still be of use. The knowledge of the incorrect prediction and the real biological results can be used to refine the model via the cycle of System Biology. In general, it could be possible to construct a model directly for the purpose of drug target prediction. For many applications, however, the construction of a biologically relevant model requires more cycles of prediction, hypotheses generation, biological experiments, and model refinements than a model constructed on the basis of diverse available biological data.

Constructing Systems Biology models The construction of a comprehensive computational model, which is the start the workflow proposed within this thesis as seen in Figure 3.1, is a demanding task. First of all, it

3.1. SEMANTIC INFORMATION IN SYSTEMS BIOLOGY

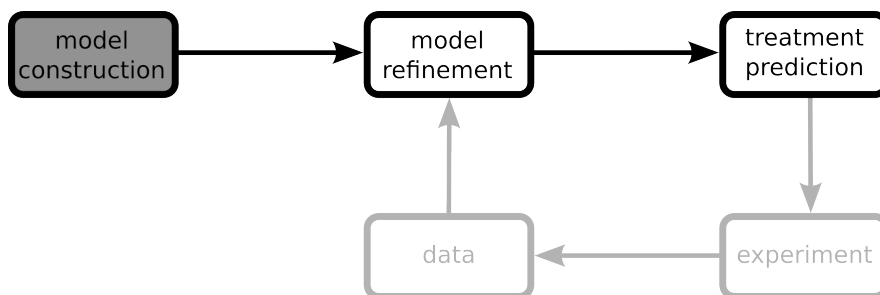


Figure 3.1: Current position in the workflow of applying Systems Biology methods to pharma research and development.

requires the integration of knowledge from various sources: text books, recent journal articles, databases, experimental results, and available Systems Biology models. This knowledge has to be collected and evaluated for relevance. Second, this vast amount of information has to be condensed into a set of preferably few mathematical equations. Finally, although much information from different sources flows into the construction of a model, not every last detail of the model might be determined. Even with lots of research one might still lack some numerical values for substance concentrations, volumes, or kinetic parameters. This can be due to the fact that this particular information can not be measured, has not been measured, yet, or simply that it has not been found in the overwhelming amount of available information.

In order to rule out the latter problem and to simplify the process of gathering relevant knowledge from the rapidly increasing amount of information, it renders necessary to develop methods for making biochemically relevant data searchable by computers. The first big step in that direction are the different recent public data repositories, which will most probably combine all knowledge in the future [Aldridge et al., 2006]. These repositories will not only be of use by storing available data in a single physical location but also by standardising the way in which this data is deposited, which simplifies automatic processing. A small number of web resources relevant to model construction is shown in Table 3.1.

A unified “language” In the construction of a kinetic model, the knowledge from various databases has to be integrated. One can for example combine the structures of certain pathways, kinetic parameters of the involved enzymes, and gene expression data under various conditions in order to construct a comprehensive mathematical model. A prerequisite for such use cases is, however, that one can compare and relate entries from various different databases such that one knows if and how one can combine

3.1. SEMANTIC INFORMATION IN SYSTEMS BIOLOGY

Table 3.1: Examples of various web repositories relevant in the construction of Systems Biology models.

| Type of data | Name | Description |
|-------------------------|--------------------------------------------------------------------------------------|------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------|
| Experimental data | ArrayExpress [Parkinson et al., 2009] | AE is a repository storing experimental data which are mainly array based. |
| | Gene Expression Omnibus [Barrett et al., 2011] | GEO is another repository containing microarray experiments and next-generation sequencing data. |
| Pathway data | PubChem BioAssays [Wang et al., 2010a] | This repository stores array descriptions and results from assays involving small compounds and siRNA. |
| | BRENDA [Scheer et al., 2011] | BRENDA contains enzyme kinetics and kinetic parameters manually collected from scientific literature. |
| | KEGG [Kanehisa et al., 2008] | KEGG encompasses various web repositories containing e.g. information on compounds, genes, and pathways. |
| Complete kinetic models | Reactome [Croft et al., 2011] | Reactome contains curated biochemical pathways and crosslinks its information to other relevant web repositories. |
| | BioModels Database [Li et al., 2010] | This is a repository for highly curated Systems Biology models from the literature. |
| | JWS Online [Olivier and Snoep, 2004] CellML repository [Lloyd et al., 2008] | JWS Online is an online simulation environment for kinetic models including a database of published models. The CellML Model Repository contains a large list of mathematical models in the CellML format. |

3.1. SEMANTIC INFORMATION IN SYSTEMS BIOLOGY

knowledge from different sources.

Of course, having a huge number of different databases at hand gives us access to lots of diverse data. Nevertheless, it comes at the price of a large number of different “naming conventions”, because different people might think of different things and still use the same word for it. A simple example for this fact is ChEBI [Degtyarenko et al., 2008] entry 4167, which is named “D-glucopyranose”. This entry is deemed equivalent to KEGG entry C00031, which is named “D-Glucose”. At the same time the ChEBI entry 17634 has the name “D-glucose”, but it is not thought to have a corresponding entry in KEGG.

For the purpose of unifying the language between different kinds of researchers web resources like Gene Ontology [Ashburner et al., 2000] have been developed. These ontologies serve a lot of purposes.

- First, they unify terms by assigning an ID, a standard name, and possible synonyms to them.
- Second, they can provide a definition of the described entity. ChEBI for example provides a chemical structure for its entries.
- Furthermore, they can provide relations between entries in their resource and entries in other resources. The types of these relations can be very diverse. ChEBI entry 4167 (“D-glucopyranose”) has an “is” relation to KEGG entry C00031 (“D-Glucose”) and is therefore supposed to describe the same chemical entity. The CGD (*Candida* Genome Database) [Skrzypek et al., 2010] entry CAL0000198 (“HXK2”) has an “is_ortholog” relation to SGD (*Saccharomyces* Genome Database) [Engel et al., 2010] entry S000003222 (“HXK2”). In turn, this gene “encodes” for the UniProt [Bairoch et al., 2009] protein P04807 (“Hexokinase-2”), whose functional classification is described by the E.C. number 2.7.1.1.
- Finally, an ontology provides relations of different types between its own entries. An example for such a relation is the ChEBI entry 4167 (“D-glucopyranose”), which is a more specific term and a direct “is_a” child of ChEBI entry 17634 (“D-glucose”). Depending on the types of used internal relations the strict term for the ontology differs. In case there are no internal relations at all, it is called a *controlled vocabulary*, if there is only one type of relation (an “is_a” relation), it is a *taxonomy*, and in case there are more types, it is an *ontology* in the strict sense.

3.1. SEMANTIC INFORMATION IN SYSTEMS BIOLOGY

Linking Systems Biology models to ontologies IDs from public ontologies do not suffer from the same ambiguities as descriptions in natural language. Therefore, they are suited much better to describe the *Biological Concept* (BC) behind a variable in a Systems Biology model or behind numerical values in a data set [Krause et al., 2011]. Furthermore, they can be processed by computers more easily.

In order to define the intended BCs behind the elements in a computational model, the MIRIAM standard [Le Novère et al., 2005] has been developed. This standard defines, how links to entries of various web resources can be written in XML using the Resource Description Framework (RDF). On top of this standard, the BioModels initiative [Laibe and Le Novère, 2007] has defined a set of useful web resources and a set of relations the annotated model elements can have to ontology entries. Up to now, MIRIAM compliant annotations have been used in many models (e.g. the models in the BioModels Database) and are supported by various software tools (e.g. semanticSBML [Schulz et al., 2006] or CoPaSi [Hoops et al., 2006]). But not only the biological aspects of a model can be assigned computer readable annotations: The Systems Biology Ontology [Le Novère, 2006] is a taxonomy defining mathematical and functional terms which can be used to describe the meaning of kinetic parameters or the role (e.g. inhibitor or catalyst) of a compound in a reaction.

Comparing Systems Biology models As already mentioned, to make full use of the available information when constructing a kinetic model one has to be able to automatically compare models and data sets. One needs to be able to search resources for appropriate models. One needs to rank the retrieved results by relevance to the current application. One needs to classify collected models and data sets according to their specific use. And one needs to be able to relate the details of models and data set, in order to see how they differ, overlap, and complement each other [Liebermeister, 2008, Krause et al., 2010]. In general, well accepted algorithms exist for all of these applications. But in order to apply them, one has to define a similarity measure for models, model elements, and data sets. And, as stated above, this measure should be computed from machine readable semantic information.

A first question one has to ask is which aspects of a Systems Biology model are supposed to be captured by the similarity measure. Complete kinetic models are always a combination of two different kinds of information: (i) the biology or “what” is described by the model (e.g. substances, proteins, or processes converting them) and (ii) the math or “how” the model is described

3.1. SEMANTIC INFORMATION IN SYSTEMS BIOLOGY

(e.g. how does the stoichiometric matrix look like or which kinetics are used).

In this chapter, a similarity measure for the “what” should be developed. Following this simple approach is advantageous from three different points of view. First, the simpler the measure the broader will be its applicability. When one neglects the mathematical aspects of the model, there is no conceptual difference between an annotated model and an annotated data set. Therefore, one will be able to compare data sets with data sets and/or models. Second, a simple measure can serve as a starting point in the development of a more complex measure incorporating the structure. Finally, sometimes the “what” of a model might already give hints on the “how”. The paper of Markevich *et al.* [Markevich et al., 2004], for example, discusses different model alternatives with different structures. Along these different structures also the biological annotations are changing, as e.g. assigned GO terms are changing in between an enzymatic reaction and its implementation in distinct reactions with mass action kinetics.

During the development of the methods introduced in the following section, a related approach to a similarity measure has been published: Henkel *et al.* [Henkel et al., 2010] have developed a method to query BioModels Database for models containing certain (or related) terms. This approach combines techniques from information retrieval with similarities defined on entries in ontologies. As the most promising measure I have developed in this thesis will follow a similar approach, I will introduce the different fields of research it is built on in the following. Apart from similarities and information retrieval, the challenges in combining knowledge from different ontologies will be addressed. Because various kinds of information from various resources might be useful for an extensive similarity measure and because established measures have been developed on the basis of individual web resources, the integration of ontologies should be regarded as a prerequisite of this measure. Furthermore, concepts of the semantic web and minimal annotations will be introduced as they give us information on the biological content of data sets and models.

3.1.2 Integrating ontologies

What should be regarded as an integrated ontology? Before the question of how one gains an integrated ontology should be answered, one first has to define what should be regarded as an integrated ontology. Or, seen from a different point of view, what will be the result of the integration process? As its main constituents, an ontology contains uniquely identifiable entries, which might be described in more or less detail, internal relations between entries, and cross-links stating how internal entries relate to entries

3.1. SEMANTIC INFORMATION IN SYSTEMS BIOLOGY

from other resources. In the scope of this thesis, I see the process of combining different ontologies as a procedure leaving us with a new, big ontology such that

- every entry from every ontology is linked to a certain entry in the integrated ontology,
- every entry in the integrated ontology is linked to at most one non-obsolete entry from a particular resource,
- every relation from every ontology has a corresponding relation in the integrated ontology, and
- cross references between two entries of two web resources should in general result in both entries being mapped to one entry in the integrated ontology.

How can ontologies be integrated? Even though it can be stated comparably simple how an integrated ontology should look like, the problem of how to integrate them practically is much more complex. This complexity arises from the facts that the different ontologies come from multiple sources, which might interpret things differently, they might describe entities to a different degree of detail, or they might simply be wrong. All of these problems can lead to inconsistencies like circles of directed relations in the integrated ontology.

These inconsistencies can be dealt with in two different ways: The methods designed for single, consistent ontologies could be modified to work on an erroneous graph or the inconsistencies have to be removed from the integrated ontology [Huang et al., 2005]. Since a consistent ontology might also be of use for other applications, the latter is more preferable.

One of the biggest problems in ontology integration has already been discussed in the literature: the problem of cycles of directed relations in the integrated graph. If one assumes that the individual ontologies do not contain these cycles, then the problem has to stem from the integration process. In order to tackle this problem, two accepted approaches repairing an integrated ontology remove cross references between the single ontologies which lead to a cycle. This is repeated until the result is a consistent ontology [Meilicke et al., 2007, Ji et al., 2009]. Another approach does not use direct cross relations as evidence to merge entries from two different web resources but takes cross references from a third ontology into account [Kirsten et al., 2007].

3.1. SEMANTIC INFORMATION IN SYSTEMS BIOLOGY

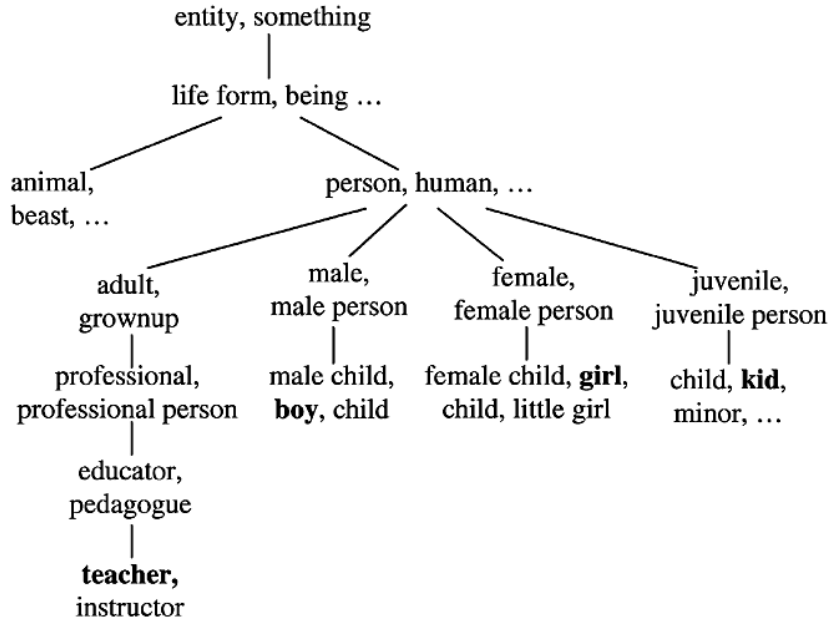


Figure 3.2: Subpart of synonyms and their specialisations as defined in WordNet [Miller et al., 1990]. Figure taken from [Li et al., 2003].

3.1.3 Similarity measure for entries in ontologies

3.1.3.1 Measures used to compare ontology entries

Prerequisites of similarity measures Similarity measures for entries in taxonomies have been discussed in the literature for decades. One of their first big applications has been the automated search for documents in large databases. Here the measures have been developed to compare words from natural language for their similarity. In order to deal with ambiguities of words the taxonomy WordNet [Miller et al., 1990] has been developed. This taxonomy divides the English language into so-called *terms* consisting of different synonyms and structures these terms as specialisations of each other in a hierarchy. A subpart of this ontology is shown in Figure 3.2.

Apart from WordNet, many approaches use a second source of information to judge the similarity of two words: their “information content”. This information content is inversely related to how often a word is used and is supposed to represent the fact that a rare term appearing in two documents is a stronger evidence of the documents’ similarity than a common term. To gain information on the frequency of words a “corpus”, which is a compilation of texts on different subjects, is used. For the English language a good corpus is the Brown Corpus of Standard American English, which contains

3.1. SEMANTIC INFORMATION IN SYSTEMS BIOLOGY

one million words extracted from texts written in 1961 [Francis and Kucera, 1964].

A third commonality between most of the early approaches is how the quality of the measures has been evaluated, or rather how much they resemble human judgement. In the nineties, Miller and Charles compiled a data set of 30 noun-noun pairs and recorded how 38 individuals rated their similarity [Miller and Charles, 1991]. The correlation between the means of their recorded similarities and the judgement of the evaluated similarity measure has henceforth been used to rate different approaches. Even though the data set has been used most often, it is actually a subset of a pair list compiled by Rubenstein and Goodenough [Rubenstein and Goodenough, 1965], which contained 65 word pairs. Nevertheless, because of its popularity the human similarities for the Miller experiment have been replicated in [Resnik, 1995].

Differences in available similarity measures Although these data sets attracted much attention, the first similarity measures for entries in an ontology have been derived beforehand. A very simple example is the concept of distance introduced by Rada *et al.* [Rada et al., 1989], which is based on the MeSH (Medical Subject Headings¹) taxonomy. This measure simply counts the number of edges in between two entries in an ontology. Despite its simplicity, this distance measure has been used in different contexts, e.g. in combination with techniques from information retrieval to improve document searches by keywords [Lee et al., 1993]. An apparent drawback of this method can be seen in the example in Figure 3.2: not all edges have the same semantic length. The edge from “entity” to “being” connects terms which are semantically more different than the terms “educator” to “teacher”. In order to deal with this problem Wang *et al.* [Wang et al., 2007] developed an approach in which the paths from both compared terms to the root are taken into account. Here, not only the number of edges appearing in both paths play a role. Also the type of the relation they represent and their distance to the compared terms is accounted for in the measure.

A common criticism of the purely ontology-based (or *edge*-based) methods has been the fact that they could not distinguish between common and rare terms. In theory, a common term appearing in two documents (probably by chance) should influence the similarity of the documents less than a rare term, which is much less likely to appear by chance. For the purpose of assessing the information content of different words various approaches use a corpus to count words in it. This information content is then combined with the structural information implicated in a taxonomy. The first approaches

¹<http://www.nlm.nih.gov/mesh>

3.1. SEMANTIC INFORMATION IN SYSTEMS BIOLOGY

using this new source of knowledge have defined the similarity of two terms as a function of the information content of their most precise superconcept [Resnik, 1995, Lin, 1998]. A superconcept is here defined as a concept on the path between the term and the root of the hierarchy. In both approaches the number of occurrences of the individual words is counted first, and then these numbers are summed up on the way upwards in the hierarchy. This ensures that more specific words are regarded as more informative as their superconcepts.

Although these so-called *node*-based methods already perform quite well, it should be clear that a combination of both types of approaches should in general work even better [Budanitsky and Hirst, 2001]. A first idea in this direction was to use information to reweigh the edges in an ontology before a distance between two terms is computed [Jiang and Conrath, 1997]. Other approaches directly combine both kinds of information independently into the similarity measure. In this setting it is possible to estimate weights, which define how much a certain type of information contributes to the similarity, in order to optimise the performance of the measure [Li et al., 2003].

3.1.3.2 Biological applications for similarity measures

In conjunction with the rise of the first biologically relevant ontologies the first similarity measures have been applied to biological problems. One of the first approaches compared sequence information from UniProt entries with the semantic similarity of their function, given by links to entries in Gene Ontology [Lord et al., 2003]. Gene Ontology terms have further been used to relate gene expression levels to the functional similarity of their products [Wang et al., 2004, Yu et al., 2007]. Based on the results of Wang *et al.* different similarity measures and the different branches of the Gene Ontology hierarchy have been evaluated to assess whether the similarities computed from them correlate well with gene expression data [Sevilla et al., 2005]. For some applications new measures have been developed, e.g. to compare the functions of all proteins expressed in an organism in order to find overlaps and differences between species [Schlicker et al., 2006] or to correlate protein families [Couto et al., 2007]. Although most of the biological applications of similarity measures are based on Gene Ontology, also other ontologies have been used. Köhler *et al.* for example used semantic information on diseases from the human phenotype ontology [Robinson et al., 2008] to determine the most probable diseases given a set of symptoms [Köhler et al., 2009].

In accordance with the huge amount of available measures and applications, also a lot of different tools, e.g. FuSSiMeG [Couto et al., 2003], FunSimMat [Schlicker and Albrecht, 2008], or G-sesame [Du et al., 2009], have

3.1. SEMANTIC INFORMATION IN SYSTEMS BIOLOGY

been developed. A comprehensive overview on similarity measures and their applications in Systems Biology can be found in Pesquita *et al.* [Pesquita *et al.*, 2009]. This review also shows that different measures are more suitable for different applications. Therefore, for every use case various measures should be evaluated.

3.1.4 Principles of information retrieval

Information retrieval (IR) is a field of research investigating the question of how to reliably find relevant documents in a large resource. As the amount and complexity of available information grows constantly, computers have to aid humans in the search for knowledge. The biggest challenge in IR is imposed by the fact that queries for documents do not need to follow a fixed format. Therefore, the question of how relevance is determined or evaluated is not determined *a priori*. Instead of simple rules, complex relations between the query and the retrieved documents are thus used to rate their relevance. These relations are then combined in heuristic similarity measures, which differ along each other in 4 central points: (i) the way how query and documents are represented, e.g. term sets or vectors [Salton, 1971], (ii) the way how terms are interrelated, e.g. not at all, based on the co-occurrences of terms [Wong *et al.*, 1987], or, as presented by Becker & Kuroopka [Becker and Kuroopka, 2003], based on the “semantic coherence” of words, (iii) the way how different terms are weighted in the similarity measure, e.g. by term frequency-inverse document frequency (TF-IDF) [Jones, 1972], which weights a term higher the more often it is used in the considered document but the less often it is used in the complete document resource, and (iv) the way how this information is combined into a similarity measure, e.g. by the cosine measure [Salton and McGill, 1986], which judges the similarity of two feature vectors by the angle between them.

3.1.5 Standards for semantic information on the internet

The annotation of data with computer readable information has not solely been invented for biological applications. Together with the introduction of various standards of the internet, the term “semantic web” has been coined. The semantic web expresses the idea to annotate documents with meta information which describe its content. This meta information might be useful for different purposes. It allows to search for relevant documents on a certain subject or, given descriptions of results from scientific literature, for the inference of new knowledge.

3.2. COMPARING MODELS AND DATA SETS BASED ON SEMANTIC INFORMATION

In order to store the meta data, the Resource Description Framework (RDF) data model has been proposed. Its main feature is the form in which information is stated: a triplet of subject, predicate, and object. For example, the sentence “glucose is a certain kind of sugar” could be expressed as a triplet (“glucose”, “is_a”, “sugar”). The single parts of the triplet are not supposed to be simple terms but references to entries from controlled vocabularies, e.g. the glucose entry in ChEBI. Given lots of elementary knowledge in this RDF format, new, more complex information can be inferred. One can for example verify that glucose really is a sugar. This information is not stated explicitly in ChEBI, but the two terms are connected via a chain of other terms and “is_a” relations, and the “is_a” relation is known to be transitive. This inference of knowledge is called reasoning.

Apart from standards on the format in which meta data on documents is stored, standards for the information placed in it are necessary. Most of these standards try to establish a set of minimal semantic information needed to judge the contents of a document. While the aforementioned MIRIAM defines this minimal information for Systems Biology models, MIAME [Brazma et al., 2001] does the same for microarray experiments. Eventually, even textual scientific information is supposed to get annotated [Cheung et al., 2010], which will make large amounts of knowledge available for automated retrieval and reasoning.

3.2 Comparing models and data sets based on semantic information

3.2.1 Combining ontologies

3.2.1.1 A library to integrate ontologies

For the purpose of relating entries from different web resources I have developed an ontology integration library called libSBAnnotation. Its development started in a project in which it has been investigated how to find overlapping elements in SBML models in order to merge them to bigger, more comprehensive models [Schulz et al., 2006, Krause et al., 2010]. Afterwards, this library has been reused to retrieve, cluster, and align similar models and data sets from large databases [Schulz et al., 2011]. It should be noted that in contrast to available ontology integration approaches the aim of the libSBAnnotation has not been the construction of a fully valid, human readable ontology. Its objective is merely to provide means for the pairwise comparison of entries from different web resources.

3.2. COMPARING MODELS AND DATA SETS BASED ON SEMANTIC INFORMATION

In the beginning, the library had a special scope. It was intended to be part of an open source tool which is suitable for researchers as well as for the industry. Because of licencing issues with some of the integrated web resources (for a full list see Table 3.2), users have been supposed to manually select the resources they need and are allowed to use. The selected resources are afterwards downloaded and integrated on their local machines, which imposed computing time and memory consumption limitations on the integration process. Since prior test implementations making use of available database systems (e.g. MySQL) have been too slow for the comparison of SBML models containing large numbers of semantic annotations and since using a database system would impose more software requirements for our tool, I have implemented the libSBAnnotation purely in Python [Van Rossum, 1995].

Being able to perform the integration process on the local machines of different users requires the libSBAnnotation to run in limited time and with less memory. One possible way to reduce the computational effort is to read single entries from ontologies one after another and integrate them on-the-fly into the libSBAnnotation. The steps of the integration and the underlying database scheme will be explained in the following paragraphs after my goals have been compared to the capabilities of other available tools.

3.2.1.2 Available frameworks incorporating semantic information

In recent years other tools sharing some similarity to the libSBAnnotation have been developed. The libAnnotationSBML [Swainston and Mendes, 2009] is a Java library acting as a wrapper for the divergent web services provided by different resources. By providing a unified interface to various web services it reduces the programming effort of integrating annotation capabilities into new tools. However, because it depends on web services it is slow when large numbers of annotations have to be compared. Another disadvantage of this approach is that it does not integrate the knowledge from its different resources in order to uncover inconsistencies in them.

BridgeDB [Van Iersel et al., 2010] is another library providing a unified interface to access various web resources. In contrast to the libAnnotationSBML it provides means to install some of the used databases locally. Nevertheless, this feature is only used to speed up the tool and not to improve the knowledge provided by individual ontologies. Further tools and libraries that provide information on the entries of various web resources exist, e.g. PICR [Côté et al., 2007b], but these are only based on single resources and therefore also do not integrate ontologies.

3.2. COMPARING MODELS AND DATA SETS BASED ON SEMANTIC INFORMATION

Table 3.2: All web repositories currently integrated during the construction of the libSBAnnotation. Web resources differ by the type of described entities (e.g. genes or organisms) and the relation types used to connect these entries.

| Web resource | Content | Relations extracted |
|--------------------------------------|----------------------------------|------------------------------------------------------------------------------------------------------------------------------------------------------------------------------|
| NCBI Taxonomy | organisms | is_a |
| Gene Ontology | compartments, processes | is_a, negatively_regulates, regulates, part_of, positively_regulates |
| ChEBI | small chemical entities (SCE) | has_functional_parent, has_parent_hydride, has_part, has_role, is_a, is_conjugate_acid_of, is_conjugate_base_of, is_enantiomer_of, is_substituent_group_from, is_tautomer_of |
| SBO | parameters | is_a |
| PubChem Compound | SCEs | |
| KEGG | SCEs, genes, reactions, enzymes, | |
| Reactome | species | |
| EntrezGene | genes | encodes, hasFunction, inOrganism, inProcess, isLocated, isPartOf |
| UniProt | proteins | encodes, hasProcess, inOrganism |
| Interpro | protein families & domains | parent, member, example, found_in |
| <i>Saccharomyces</i> Genome Database | genes | |
| <i>Candida</i> Genome Database | genes | |

3.2.1.3 Database schema of the libSBAnnotation

Although the libSBAnnotation has been developed exclusively in Python, I have tested how relational database systems would perform in various applications. For this purpose, I have developed the database schema shown in Figure 3.3, which has later been used as the basis of the Python implementation.

The schema is centred around the AbstractItem table. This table con-

3.2. COMPARING MODELS AND DATA SETS BASED ON SEMANTIC INFORMATION

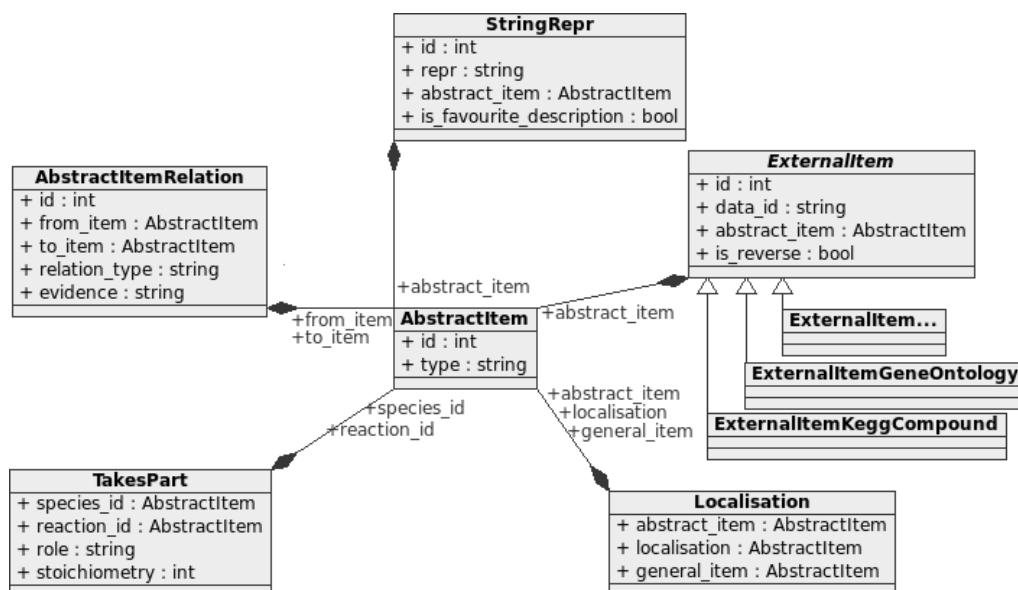


Figure 3.3: UML diagram of the database scheme underlying the libSBAnnotation.

tains volatile IDs for so-called *Biological Concepts* (BCs) and the type of information described by this concept. BCs can describe various kinds of entities in my implementation, e.g. proteins, compartments, or processes. Various synonyms as well as a preferred term for the BCs are stored in the StringRepr table.

In order to relate BCs to entries from other web resources the ExternalItem* tables have been created. Using separate tables for all referenced resources has the disadvantage that new tables have to be created each time new web resources are referred to. Nevertheless, the Python implementation is faster for many relevant queries and the abstract_item entry is unique among every ExternalItem* table. Thus, it can be used as a *key* for faster access to a specific column and to easily check for consistencies in the inserted data.

The different kinds of relations which can be used to relate BCs are stored in 3 different tables: (i) Localisation is a table for special localised compounds from Reactome, (ii) TakesPart is a table storing the stoichiometry of reactions, and (iii) AbstractItemRelations is a table containing all other types of binary relations in which two AIs are involved. The latter table contains all relations directly extracted from various web resources (see Table 3.2 for a comprehensive list) together with the name of the resource from which it has been extracted.

3.2. COMPARING MODELS AND DATA SETS BASED ON SEMANTIC INFORMATION

3.2.1.4 Integrating information from different ontologies

As previously discussed, the integration process has to be done on-the-fly due to constraints on time and memory usage. Therefore, the entries of the different web resources are included one after another. This means that in each step one has to consider how to integrate a single entry or a single relation into the constantly growing ontology. Depending on whether this information is already included in the ontology it has to be merged with existing knowledge or it can simply be added to it.

In order to simplify the integration process I avoid merging relations. This is done by neglecting sources of information which provide identical relations between the same types of BCs. The nodes, however, can be taken from different sources and can therefore contradict each other. By far the most common problem of this kind arises when one web resource states that two of its entries a_1 and a_2 are identical with an entry b_1 from a different resource; or stating it differently, when a node $\{a_1, b_1\}$ has to be merged with a new node $\{a_2, b_1\}$. In such cases two new nodes $\{a_1\}$ and $\{a_2\}$ are constructed and connected to the initial node $\{b_1\}$ by *is_a* relations. Existing relations of the prior node $\{a_1, b_1\}$ which had been extracted from the a resource are then moved to the $\{a_1\}$ node.

A problem which frequently arises from this operation (and from possible inconsistencies between the resources) are cycles of directed relations. These cycles are found in a post-processing step and repaired by removing one of the relations (preferably one which had been inserted automatically) in it.

Using these simple rules one is able to construct an ontology which is consistent according to my definitions. It should be noted that this ontology might not be of the same quality as a manually curated one. Nevertheless, it provides good means to compare pairs of entries from various databases.

3.2.2 Similarity measures for Biological Concepts

The similarity measures presented in this chapter are built in a modular manner. The similarity of complete models is dependent on the similarity of the models' elements which is in turn dependent on the similarity of annotations and single BCs. Because of this modularity the similarity of BCs is discussed first, before different model similarity measures are introduced.

3.2.2.1 Li's modular similarity measure

According to Pesquita *et al.* [Pesquita et al., 2009] different measures are more appropriate for different applications. Thus, researchers should not

3.2. COMPARING MODELS AND DATA SETS BASED ON SEMANTIC INFORMATION

rely on a single measure to produce appropriate results but they should rather compare the applicability of different approaches. Li *et al.* proposed in [Li et al., 2003] a set of similarity measures built out of three parametrised components. Each of these components uses a different kind of information to judge the similarity of two ontology entries. My idea has therefore been to test the individual components on whether they improve the quality of the introduced similarity measures. Given the case that an appropriate amount of test data is available, one could even try to estimate parameters in the individual components of the similarity measures.

Mathematical notation Before I discuss Li’s measure in detail, I first have to introduce some formal notation for models, model elements, their annotations, the referenced resource entries or BCs, and the relations between them, which will be used throughout this thesis. The rest of this methods section is based on [Schulz et al., 2011]. For more details the reader is referred to this publication.

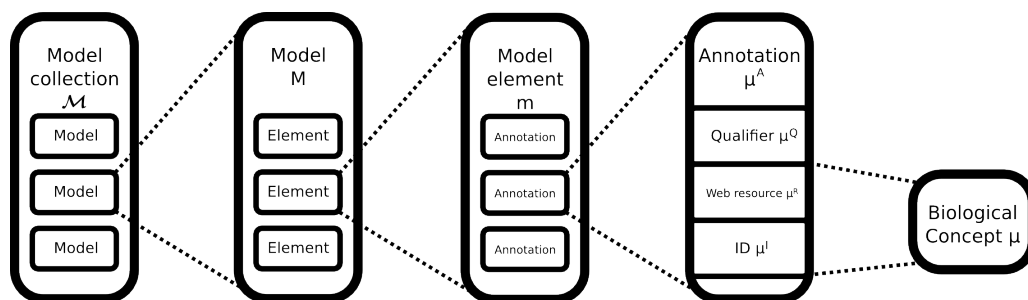


Figure 3.4: Subdivision of a model collection into single annotations and their related Biological Concepts.

Formally, I regard annotations, elements, and models as nested sets (compare Figure 3.4). MIRIAM-compliant annotations are regarded as parts of elements $\mu^A \in m$, which are a part of a model $m \in M$, which are a part of a model list $M \in \mathcal{M}$. Depending on the measure I will discuss in the following, the layer of the model elements can also be disregarded. In this case the annotations are directly a part of the model. Each annotation μ^A relates the model element to an identifier (ID) μ^I from a web resource μ^R , while the qualifier μ^Q specifies the relation between the element and the corresponding resource entry. Thus, an annotation is formally a triple $\mu^A = (\mu^R, \mu^I, \mu^Q)$. Given the knowledge from the libSBAnnotation, all web resource entries (μ^R, μ^I) can be mapped to Biological Concepts μ . Since I am not interested in the fact which specific resource was used to annotate a model element, I can internally represent an annotation by the tuple $\mu^A = (\mu, \mu^Q)$.

3.2. COMPARING MODELS AND DATA SETS BASED ON SEMANTIC INFORMATION

In the libSBAnnotation, relation edges connecting the BCs μ and ν can be described as triples $r = r(\mu, \nu, relation_type)$ (for a list of possible relation types, see Table A.1 in the Appendix). I denote the set of outgoing relation edges from a BC $R(\mu)$ and the set of outgoing relation edges of a certain type is called $R(\mu, relation_type)$. Be the depth $d(\mu)$ of a BC defined as the number of “is_a” relations to a root in the ontology and the height $h(\mu)$ by the maximal number of “is_a” relations to any of the leaves below it. Finally, I define the frequency of a BC in BioModels Database as c_μ , the cumulative frequency as $c_\mu^c = c_\mu + \sum_{\xi:r(\mu,\xi,t)\in R(\mu,is_a)} c_\xi^c$ which sums up the frequencies of all its children in the “is_a” hierarchy, and the total number of annotation appearances including pseudocounts as $c_\Omega = 1 + \sum_{\forall \xi:c_\xi > 0} (c_\xi + 1)$. To include MIRIAM annotations that do not match any known BC in the considered ontology, one can create new BCs without any relations to other concepts and then treat them as if they had already been contained in the ontology.

Li’s measure Li’s similarity measure takes two different sources of knowledge into account: a taxonomy of a natural language and a text corpus with which the information content of terms is judged. It is built from three independent factors $f_{1/2/3}$ which measure the similarity of two words: f_1 takes the number of relations l on the shortest path between two terms into account, f_2 contains a factor incorporating the depth of the lowest common ancestor of two terms in the ontology, and f_3 is computed from the information content of the lowest common ancestor, which results from the composition of the text corpus. From these factors, Li *et al.* set up their similarity measure

$$\sigma(\mu, \nu) = f_1(\mu, \nu) \cdot f_2(\mu, \nu) \cdot f_3(\mu, \nu).$$

Transferring this measure to my ontology leaves me with two problems. First, I am not computing the similarity from a taxonomy but from an ontology. Therefore, different relation types have to be taken into account when computing f_1 . Second, the entries which are compared in my ontology are not necessarily members of the same “is_a” relation hierarchy, e.g. when genes and proteins are compared. This means that few pairs of compared entries have a unique lowest common ancestor. Therefore, the $f_{2/3}$ factors should be modified such that they incorporate the depth and the information of the compared entries, which are always available.

Adaption of Li’s measure to compare entries from full ontologies
In order to deal with the aforementioned problems I have adapted Li’s similarity

$$\sigma_{BC}^{Li}(\mu, \nu) = f_1(\mu, \nu) \cdot f_2^{Li}(\mu, \nu) \cdot f_3^{Li}(\mu, \nu)$$

3.2. COMPARING MODELS AND DATA SETS BASED ON SEMANTIC INFORMATION

by modifying the single factors.

The ontology distance factor f_1 is defined as the product of scores for the individual relations on a path between two entries. In case different paths between two entries exist, it is defined as the maximum over these paths. Numerical values $f_{\text{rts}}(t)$ for the different relation types t are in the range between 0 and 1; heuristically determined values are given in Table A.1 in the Appendix. The factor f_1 can be recursively defined as

$$\begin{aligned} f_1(\mu, \mu) &= 1 \\ f_1(\mu, \nu) &= \max_{t, \xi: r(\mu, \xi, t) \in R(\mu)} (f_{\text{rts}}(t) \cdot f_1(\xi, \nu)). \end{aligned} \quad (3.1)$$

In case that there exists no relation path between μ and ν , f_1 is set to 0.

Depending on the specificity of two entries, the length of the path between them should be reweighed. For more specific entries a certain path length should result in a higher similarity than the same path length between un-specific entries. The ontology depth factor f_2 incorporates this specificity by measuring the *relative* depth of two entries in their ontology branches:

$$f_2^{\text{Li}}(\mu, \nu) = \tanh \left(\frac{3}{2} \left(\frac{d(\mu) + 1}{d(\mu) + h(\mu) + 1} + \frac{d(\nu) + 1}{d(\nu) + h(\nu) + 1} \right) \right). \quad (3.2)$$

The prefactor 3/2 here has been chosen ad-hoc to use the nonlinear range of the hyperbolic tangent function.

To deal with the fact that Biological Concepts appear in Systems Biology models with different frequencies and that they therefore contribute differently to a model’s “identity”, the information content factor f_3 is included into the similarity measure:

$$f_3^{\text{Li}}(\mu, \nu) = \tanh \left(-\log \left(\frac{\min(c_\mu^c, c_\nu^c)}{c_\Omega} \right) \right). \quad (3.3)$$

This factor decreases the similarity of common BCs such as ATP and thus decreases their weight in the comparison of complete models.

A general problem I see in the definition of the measure is the independence assumption between the terms f_1 and f_2 . In my opinion this independence assumption is not justified. Given a simple, balanced taxonomy of “is_a” relations and two entries having the maximal distance (measured in length of the path between them) in it, it is evident that each of them also has to have maximal depth. Furthermore, given that two entries have minimal depth, they both have to be the root and have distance 0. Thus, I propose to drop the independence assumption and modify the way in which both factors contribute to the similarity.

3.2. COMPARING MODELS AND DATA SETS BASED ON SEMANTIC INFORMATION

3.2.2.2 A modular measure for comparing Biological Concepts

A modified version of Li’s similarity measure assuming a dependence between the distance and the depth factor is given by

$$\sigma_{\text{BC}}^{\text{DD}}(\mu, \nu) = f_1(\mu, \nu)^{f_2(\mu, \nu)} \cdot f_3(\mu) \cdot f_3(\nu) \quad (3.4)$$

where

$$f_2(\mu, \nu) = \frac{2}{d(\mu) + d(\nu) + 2}$$
$$f_3(\mu) = 1 - \frac{\log(c_\mu^c + 1)}{\log c_\Omega}.$$

This formula has been designed respecting the idea that the conceptual distance between two entries connected by a relation declines exponentially with their depth in the ontology. Therefore, a path length l of entries in depth d should result in the same similarity as a path length $2l$ in depth $2d$.

The change in how f_3 is implemented in Equation 3.4 results from the fact that the information content of both entries should be regarded as independent. Furthermore, the normalisation using the tangens hyperbolicus seemed arbitrary and was replaced by a more simple linear normalisation to $\log c_\Omega$.

3.2.3 Similarity measures for annotated data sets and models

3.2.3.1 Comparing MIRIAM annotations

As shown in Figure 3.4, the difference between a MIRIAM compliant annotation and a Biological Concept is the qualifier. This qualifier states the relation between the model element and the web resource entry. In order to compare annotations, I complement the similarity measure with a factor accounting for the different qualifiers

$$\sigma_{\text{An}}(\mu^{\text{A}}, \nu^{\text{A}}) = f_{\text{qsm}}(\mu^{\text{Q}}, \nu^{\text{Q}}) \cdot \sigma_{\text{BC}}(\mu, \nu). \quad (3.5)$$

This factor is independent of the similarity of the BCs and is supposed to decrease with the distance the qualifier implies (see Table A.2 in the Appendix for numerical values). Nevertheless, both qualifiers are not regarded as independent because certain combinations of qualifiers can imply a closer relation of the model elements. An example for such a relation is a protein which is annotated with the gene it is encoded by. In case this protein is compared to itself, the combination of two “isEncodedBy” qualifiers should increase the similarity.

3.2. COMPARING MODELS AND DATA SETS BASED ON SEMANTIC INFORMATION

3.2.3.2 The preference-based measure

Comparing model elements In SBML, the biological meaning of model elements can be expressed by a set of annotations. These annotations may point to identical Biological Concepts in different web resources but they may as well express different aspects of the element. A protein for example, could carry an annotation which uses the “is” qualifier to point to a UniProt entry and another annotation stating that this protein “isEncodedBy” a gene from the *Saccharomyces* Genome Database.

In order to deal with the different ways in which a model element can be annotated, I have developed similarity measures for elements and complete models. The idea behind this measure, which is similar to the one proposed in [Köhler et al., 2009], is that for every annotation in one of the compared elements one tries to find the closest annotation in the other element. Because of the fact that annotations can “choose” a matching annotation from the other element, I call the similarity measure

$$\sigma_{\text{El}}^{\text{Pref}}(m, n) = \frac{\sum_{\mu^A \in m} \max_{\nu^A \in n} \sigma_{\text{An}}(\mu^A, \nu^A) + \sum_{\nu^A \in n} \max_{\mu^A \in m} \sigma_{\text{An}}(\mu^A, \nu^A)}{|m| + |n|},$$

a preference-based measure. In this formula $|m|$ denotes the number of annotations assigned to model element m . Given the case that one of the elements is not annotated, one of the maxima would not be defined. Therefore, I set the similarity to a value $\varepsilon_{\text{El}} \geq 0$, when one element has no annotations. This value represents the small probability that randomly picked elements have the same meaning, because no annotations supporting their dissimilarity are known.

Comparing models The preference-based similarity measure for models

$$\sigma_{\text{Mo}}^{\text{Pref}}(M, N) = \frac{\sum_{m \in M} \max_{n \in N} \sigma_{\text{El}}(m, n) + \sum_{n \in N} \max_{m \in M} \sigma_{\text{El}}(m, n)}{|M| + |N|} \quad (3.6)$$

follows the same reasoning as the aforementioned measure for elements and is computed from a similar formula.

A strange inconsistency of this measure can be seen in Figure 3.5. The similarity of the two models M and N is computed as $\frac{3+\varepsilon}{4}$ while the similarity of N with itself has a value of $\frac{1+\varepsilon}{2}$. Thus, N is more similar to M than to itself. To correct for this behaviour, which results from elements lacking annotations a normalised similarity

$$\hat{\sigma}_{\text{Mo}}^{\text{Pref}}(M, N) = \frac{\sigma_{\text{Mo}}^{\text{Pref}}(M, N)}{\max(\sigma_{\text{Mo}}^{\text{Pref}}(M, M), \sigma_{\text{Mo}}^{\text{Pref}}(N, N))} \quad (3.7)$$

3.2. COMPARING MODELS AND DATA SETS BASED ON SEMANTIC INFORMATION

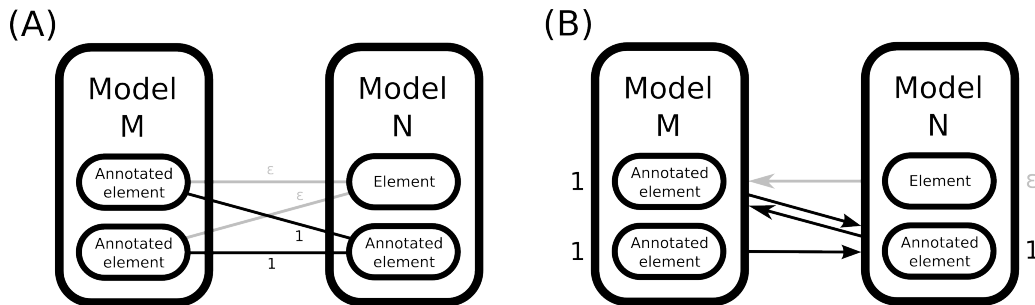


Figure 3.5: Preference-based similarity measure. (A) Pairwise element similarities are depicted for the pairs of model elements. (B) Each element is assigned the maximum of the similarities it is involved in. The similarity of the compared models is now $\sigma_{\text{Mo}}^{\text{Pref}}(M, N) = \frac{1+1+\epsilon+1}{2+2} = \frac{3+\epsilon}{4}$.

could be used. Using this formula N would, as expected, be more similar to itself than to M . Nevertheless, this formula leads to other problems as two models sharing one single annotation and containing the same number of not annotated elements would always have a similarity of 1.

3.2.3.3 Vector space model

The vector space model (VSM) [Salton, 1971] (for more details see [van Rijsbergen, 1979, Berry et al., 1999]) follows the idea to represent documents by vectors. These vectors reside in the space of so-called features i and their numerical coefficients $a_{i,d}$ denote if and how a document d is referring to each feature:

$$v_d = \begin{pmatrix} a_{1,d} \\ a_{2,d} \\ \vdots \end{pmatrix}.$$

Instead of regarding v_d as a vector in the space of features t_i it is also possible to include explicit knowledge about the features and rewrite v_d as a linear combination

$$v_d = \sum_i a_{i,d} t_i.$$

In the most simple case, one assumes independence of the features and defines their associated vectors to be unit vectors ($t_i = e_i$).

For the application of comparing Systems Biology models using the VSM I disregard the layer of model elements and describe the models by vectors in the space of Biological Concepts. Each model M is assigned a vector v_M which contains a one ($v_{iM} = 1$) for all BCs i which are referred to by

3.2. COMPARING MODELS AND DATA SETS BASED ON SEMANTIC INFORMATION

the model. All other coefficients are set to 0. In principle it is possible to incorporate the qualifiers of the annotations to assign the coefficient v_{iM} values from the range $[0, 1]$ (e.g. $v_{iM} = \sqrt{f_{\text{qsm}}(\mu^{\text{Q}}, \mu^{\text{Q}})}$), but for simplicity this is not investigated in the scope of this thesis.

Cosine similarity measure Different measures have been introduced which can be used to determine the similarity of feature vectors (e.g. Dice’s coefficient [Dice, 1945] or the overlap coefficient). A measure with a comprehensible geometrical interpretation is the cosine of the angle between the two compared vectors [Salton and McGill, 1986], which is given by the formula

$$\sigma(M, N) = \frac{v_M^T \cdot v_N}{\|v_M\|_2 \|v_N\|_2}.$$

Transforming the space of features This simple measure already leads to convincing results. Nevertheless, it has the problem that if slightly different BCs are used to annotate elements of two models (probable if models are annotated by different people), this simple measure would not capture any similarity. To deal with this problem I apply a transformation A to the basis vectors of the feature space. This transformation is supposed to reduce the angle between similar features to make them lose their orthogonality and is based on the idea of the Topic-based vector space model (TVSM) [Becker and Kurovka, 2003]:

$$\sigma(M, N) = \frac{v_M^T A^T A v_N}{\sqrt{v_M^T A^T A v_M} \sqrt{v_N^T A^T A v_N}}.$$

In the following, I will discuss how this transformation or rather how the matrix $S = A^T A$ can be constructed. For this purpose I compute the similarity of two models M_1 and N_1 which each contain only one annotation (w.l.o.g. the i^{th} and j^{th} feature representing the BCs μ and ν). This similarity then computes as

$$\begin{aligned} \sigma_{\text{Mo}}^{\text{TVSM}}(M_1, N_1) &= \frac{v_M^T S v_N}{\sqrt{v_M^T S v_M} \sqrt{v_N^T S v_N}} & (3.8) \\ &= \frac{e_i^T S e_j}{\sqrt{e_i^T S e_i} \sqrt{e_j^T S e_j}} \\ &= \frac{S_{ij}}{\sqrt{S_{ii}} \sqrt{S_{jj}}}. \end{aligned}$$

3.3. RETRIEVAL, ALIGNMENT, AND CLUSTERING OF MODELS AND DATA SETS

Given the idea that the similarity of the models M_1 and N_1 should be equal to the similarity of the BCs μ and ν , it is evident that the matrix S is made out of normalised similarities of the Biological Concepts. If one requires the similarity measure σ_{An} to fulfil $\sigma_{\text{An}}(\mu, \mu) = 1$, one ends up with $\sigma(M_1, N_1) = S_{ij} \stackrel{!}{=} \sigma_{\text{An}}(\mu, \nu)$ and the pairwise BC similarities can directly be inserted into the matrix S .

3.3 Retrieval, alignment, and clustering of models and data sets

3.3.1 Assessing the quality of different similarity measures

In the following I will evaluate the quality of different similarity measures by using them to cluster models from the BioModels database or compare clusters of them. For this purpose, I have compiled two sets of models, a small set of 14 well-known models on 4 different pathways (see Table 3.3) and a large set for which the division into predefined clusters has been done semi-automatically according to the pathway annotations on their “model” elements (see Supplementary Table A.3).

Table 3.3: Small set of models from the BioModels database, which have been grouped manually according to the pathways they describe. The right column does not show the complete IDs from the BioModels Database but only its significant parts. Full IDs consist of the string “BIOMD” followed by 10 digits.

| Described pathway | BioModels Database model identifier |
|--------------------|-------------------------------------|
| Glycolysis | 70, 71, 211 |
| Circadian clock | 16, 21, 22 |
| Cell cycle | 5, 7, 8, 111 |
| MAP kinase cascade | 26, 27, 28, 29 |

Silhouette coefficient My first idea to evaluate how similar the models in a predefined group are, is to compare intra- and inter-group similarity by the silhouette coefficient [Kaufman and Rousseeuw, 1990]

$$\text{sc}(\mathcal{M}) = \frac{\sum_{M \in \mathcal{M}} \frac{\iota(M) - \epsilon(M)}{\max(\iota(M), \epsilon(M))}}{|\mathcal{M}|}, \quad (3.9)$$

3.3. RETRIEVAL, ALIGNMENT, AND CLUSTERING OF MODELS AND DATA SETS

where

$$\begin{aligned}\epsilon(M) &= \max_{\mathcal{C} \in \mathbf{C}, M \notin \mathcal{C}} \frac{\sum_{N \in \mathcal{C}} \sigma_{\text{Mo}}(M, N)}{|\mathcal{C}|}, \\ \iota(M) &= \max_{\mathcal{C} \in \mathbf{C}, M \in \mathcal{C}} \frac{\sum_{N \in \mathcal{C}, N \neq M} \sigma_{\text{Mo}}(M, N)}{|\mathcal{C}|},\end{aligned}\tag{3.10}$$

\mathcal{M} is the set of benchmark models, and \mathbf{C} is the set of predefined biological model groups. Given a high silhouette coefficient, one knows that the similarities within a group (ι) are higher than similarities between two different groups (ϵ). This value therefore measures how clearly models from different groups are separated.

Jaccard coefficient My second idea to compare measures is to evaluate clusterings of the models. For this purpose I use the different similarity measures to perform an agglomerative clustering of the models with average linkage. This is continued until the results contains as many clusters as there are predefined groups in the data set. These clusters are then compared to the groups by the Jaccard similarity coefficient [Jaccard, 1901]

$$\text{jac} = \frac{O_{11}}{O_{01} + O_{10} + O_{11}},\tag{3.11}$$

where O_{11} is the number of model pairs sharing the same group and the same cluster and O_{10} and O_{01} are the number of pairs appearing exclusively either in the same group or the same cluster. This value is supposed to determine how closely the clustering based on the compared measures is related to the predefined groups.

3.3.2 Evaluation of the different measures

Which is the best model similarity measure? The results of the comprehensive analyses of the different similarity measures evaluated by different coefficients and using different data sets are shown in Table 3.4. In general, no single measure performs best with respect to all four tests. Furthermore, for most tests the values of the quality measures are quite similar. The only test showing a distinctively superior measure is the Jaccard coefficient for the clustering of the large model set. Here the TVSM measure which disregards the information content of Biological Concepts ($f_3 = 1$) is clearly the best. Since this measure can be evaluated much faster and since it is also applicable to annotated data sets or any other type of information that can be transferred to a list of BCs, I have used this measure in the web tool we

3.3. RETRIEVAL, ALIGNMENT, AND CLUSTERING OF MODELS AND DATA SETS

Table 3.4: Evaluation of model similarity measures with predefined model groups. Different variants of normalised similarity measures (rows) were compared for the small and large model benchmark sets. The silhouette coefficient scores the similarities of models within and between groups. For computing the Jaccard coefficient, the models were clustered by agglomerative clustering with average linkage and with the respective similarity measure. The dendrograms were cut at a height where the numbers of clusters and predefined model groups were identical (4 groups for the small benchmark set; 34 for the large benchmark set). The silhouette and the Jaccard coefficient assume values in the range between -1 and 1 or 0 and 1, respectively, with higher values denoting a better classification.

| Similarity measure $\hat{\sigma}$ (normalised) | Silhouette coef. | | Jaccard coef. | |
|-------------------------------------------------------------------|------------------|-----------|---------------|-----------|
| | Small set | Large set | Small set | Large set |
| TVSM, S given by σ_{BC}^{DD} | .657 | .0982 | 1 | .284 |
| with $S = f_1^{f_2}$ | .705 | .136 | 1 | .377 |
| with $S = I$ | .766 | .146 | 1 | .356 |
| Preference-based with σ_{El}^{Pref} and σ_{BC}^{Li} | .746 | .108 | 1 | .231 |
| setting $c_\mu^c = c_\mu$ | .746 | .108 | 1 | .231 |
| setting $f_3^{Li} = 1$ | .746 | .104 | 1 | .229 |
| Preference-based with σ_{El}^{Pref} and σ_{BC}^{DD} | .741 | .118 | 1 | .231 |
| setting $c_\mu^c = c_\mu$ | .738 | .120 | 1 | .231 |
| setting $f_3 = 1$ | .700 | .123 | 1 | .269 |
| additionally setting $f_2 = 1$ | .720 | .122 | 1 | .254 |
| additionally without libSBAnnotation | .746 | .101 | 1 | .229 |
| additionally setting $f_{qsm} = 1$ | .709 | .095 | .556 | .262 |

have developed in our group (<http://semanticsbml.org>). The preference-based measures also performed well, but their additional complexity is not compensated for by better results.

Which is the best similarity measure for Biological Concepts? A second question I want to answer with this large scale analysis is which kind of measure should be used to compare BCs, σ_{BC}^{Li} or σ_{BC}^{DD} . Also this point cannot be answered with certainty from the results in Table 3.4. If one again puts more emphasis on the Jaccard coefficient for the large data set, then σ_{BC}^{DD} (especially setting $f_3 = 1$) shows much better results than the measures using σ_{BC}^{Li} . Since I also suppose the measure to be more properly justified,

3.3. RETRIEVAL, ALIGNMENT, AND CLUSTERING OF MODELS AND DATA SETS

as it does not assume independence of term distance and term specificity, I prefer σ_{BC}^{DD} for all further considerations.

The final point addressed by the analysis was whether and how different kinds of information (distance, depth, and information) on compared BCs should affect their similarity. In this point the data was not clear either. The only change introduced to the compared measures which has improved results under more than one condition is the neglect of the information factor f_3 . Thus, I suppose that the information of a BC should not affect how it is contributing to models' similarities. This finding is in agreement with the results presented in [Li et al., 2003]. Disregarding f_3 does, however, not mean that one forgets the knowledge in the "information" content of BCs. This knowledge can be taken into account in the computation of p-values as shown in section B.1 in the Appendix.

Evaluation of the test data A final point on the data in Table 3.4 is the general decline in the performance of the measures in between the small and the large data set. The major reason for this might be the fact that the definition of the model groups is not perfect. One example for this is the distinction between the first and the fifth group in Supplementary Table A.3. These models cannot be distinguished based on semantic annotations as they contain too few of them. Nevertheless, I did not change the predefined groups. These groups have been assigned based on their model annotations, which are not considered by the similarity measures, and modifying the groups based on the results I have gained so far would bias the outcome towards too positive results.

Optimising parameters of the measures In principle, one of these tests could be used to optimise the different parameters used in the similarity measures. But since too few data is available I have been unable to reliably estimate the large number of parameters in most similarity measures. When running a parameter optimisation one ends up with strongly biased results. An example of this is that pairs of "isVersionOf" qualifiers tend to get higher scores than "is" pairs, because they are used in many models describing proteins involved in signalling. In general, one can assume that a bias towards the usage of certain qualifiers and BCs is introduced by the fact that only a few people are annotating BioModels. Therefore, I propose for the estimation step to be repeated as soon as models from more sources and a gold standard of model categories are available.

However, not having optimised the parameters in my measures does not necessarily alter the quality of the similarities. The similarity values and the

3.3. RETRIEVAL, ALIGNMENT, AND CLUSTERING OF MODELS AND DATA SETS

ranking of retrieved models according to the similarities are not very sensitive to variations in the parameter values (compare Supplementary Figure A.4).

3.3.3 Matching experimental data to models

3.3.3.1 Oscillating yeast genes

The vector-based similarity measures have multiple benefits. One of them is their simplicity, which enables their use in diverse applications. Beyond the comparison of models it enables us to relate data sets with themselves and with models. This application can be useful when experimental data is needed to which a model with poorly determined parameters can be fitted. Furthermore, it enables the retrieval of models by an experimental data set. The retrieved models could then serve as a basis for further modelling efforts. In order to demonstrate the applicability of the TVSM measure to compare models with data sets, I select an experimental data set from the literature and search for models in the BioModels Database covering parts of the data.

For this purpose I use a data set describing oscillating genes, which are coupled to bursts in DNA replication in yeast. The data of Klevecz *et al.* [Klevecz *et al.*, 2004] show that these genes are part of sulphur and methionine metabolism pathways and involved in proteolysis, the ribosomes, and the DNA polymerase.

I use the tool Annotate Your Model (AYM, <http://semanticsbml.org/aym>) to automatically annotate the list of gene names from Klevecz *et al.* and search for similarly annotated models (the process is described in detail online <http://semanticsbml.org/aym/default/examples>).

The results of this retrieval process are shown in Figure 3.6. Retrieved models describe amino acid metabolism (BioModels with IDs ending in 66, 68, 90, 190, and 212), include ubiquitination steps (105, 154–159, 186, 187, and 293), or contain a DNA polymerisation reaction (15). The functional categories of the pathways identified by Klevecz *et al.* and the pathways described by the retrieved models therefore overlap quite well. Nevertheless, the number of annotations co-occurring in the data set and in any of the models is comparably small.

For future applications such a retrieval step might become valuable in the the analysis of data sets, given that enough knowledge is available in the form of annotated models.

3.3. RETRIEVAL, ALIGNMENT, AND CLUSTERING OF MODELS AND DATA SETS
























| Model | BioModel | Similarity | p-Value | Overlap | p-Value | |
|------------------------------------------------------|-----------------|------------|---------|---------|---------|-------------------------------------------------------------------------------------|
| Wolf2001_respiratory_oscillations | BIOMD0000000090 | 0.207 | <=1e-3 | 6 | 6.6e-09 |  |
| Chassagnole2001_Threonine_Synthesis | BIOMD0000000066 | 0.184 | <=1e-3 | 4 | 1.5e-05 |  |
| Curien2009_Aspartate_Metabolism | BIOMD0000000212 | 0.170 | <=1e-3 | 5 | 3.6e-07 |  |
| Curien2003_MetThr_synthesis | BIOMD0000000068 | 0.141 | <=1e-3 | 2 | 1.0e-02 |  |
| Proctor2007_ubiquitine | BIOMD0000000105 | 0.098 | 2.0e-03 | 1 | 1.4e-01 |  |
| Curto1998_purineMetabol | BIOMD0000000015 | 0.063 | 1.1e-02 | 2 | 1.0e-02 |  |
| Ibrahim2008_Spindle_Assembly_Checkpoint_dissociation | BIOMD0000000186 | 0.057 | 1.8e-02 | 0 | 1.0e+00 |  |
| Ibrahim2008_Spindle_Assembly_Checkpoint_convey | BIOMD0000000187 | 0.057 | 1.8e-02 | 0 | 1.0e+00 |  |
| Rodriguez-Caso2006_Polyamine_Metabolism | BIOMD0000000190 | 0.040 | 7.1e-02 | 1 | 1.4e-01 |  |
| Nijhout2004_Folate_Cycle | BIOMD0000000213 | 0.032 | 1.1e-01 | 1 | 1.4e-01 |  |
| Morrison1989_FolateCycle | BIOMD0000000018 | 0.030 | 1.3e-01 | 1 | 1.4e-01 |  |
| Zatorsky2006_p53_Model3 | BIOMD0000000154 | 0.023 | 2.5e-01 | 0 | 1.0e+00 |  |
| Zatorsky2006_p53_Model6 | BIOMD0000000155 | 0.023 | 2.5e-01 | 0 | 1.0e+00 |  |
| Hunziker2010_p53_StressSpecificResponse | BIOMD0000000252 | 0.023 | 2.5e-01 | 0 | 1.0e+00 |  |
| Zatorsky2006_p53_Model5 | BIOMD0000000156 | 0.022 | 2.7e-01 | 0 | 1.0e+00 |  |
| Zatorsky2006_p53_Model4 | BIOMD0000000157 | 0.022 | 2.7e-01 | 0 | 1.0e+00 |  |
| Zatorsky2006_p53_Model2 | BIOMD0000000158 | 0.022 | 2.7e-01 | 0 | 1.0e+00 |  |
| Zatorsky2006_p53_Model1 | BIOMD0000000159 | 0.022 | 2.7e-01 | 0 | 1.0e+00 |  |
| Proctor2008_p53_Mdm2_ATM | BIOMD0000000188 | 0.013 | 4.3e-01 | 0 | 1.0e+00 |  |
| McClean2007_CrossTalk | BIOMD0000000116 | 0.012 | 4.7e-01 | 0 | 1.0e+00 |  |
| Proctor2008_p53_Mdm2_ARF | BIOMD0000000189 | 0.012 | 4.9e-01 | 0 | 1.0e+00 |  |
| Haberichter2007_cellcycle | BIOMD0000000109 | 0.011 | 5.0e-01 | 0 | 1.0e+00 |  |
| Sasagawa2005_MAPK | BIOMD0000000049 | 0.006 | 5.5e-01 | 0 | 1.0e+00 |  |

Figure 3.6: List of retrieved BioModels when querying the database with the data set of Klevecz *et al.*. The pathways of the described models partially overlap with the pathways in which the genes from the data set participate in.

3.3.3.2 Arachidonic acid pathway

One of the running examples throughout this work is the determination of drug targets in the arachidonic acid pathway. In order to find relevant models of this pathway that are already available in the BioModels Database, I start a model retrieval using only the annotation for arachidonic acid. For this purpose I create a single element in a data set using the AYM website, annotate it with the ChEBI entry for “arachidonic acid” (CHEBI:15843), and retrieve models similar to this data set.

This very specific query only results in one single model, the arachidonic acid pathway of Yang *et al.* [Yang and Sze, 2007] (BioModel 106). It shows a similarity of 0.141 with my query data set, which is highly significant (p-value $\leq 10^{-3}$), and an overlap of 1, which is significant as well (p-value = $6.1 \cdot 10^{-3}$).

In a next step, I use this model to search for other models being able to extend it. This retrieval reveals no models having a significant overlap. Even though the overlap of the retrieved results is bigger than or equal to the overlap between my initial data set and model 106, it is less significant (p-value ≥ 0.18). The reason for this lack in significance is the fact that the Yang model also contains very common annotations (e.g. the Gene Ontology term for “cell”), which makes other models more likely to share a few annotations with it.

Based on these results, only one single relevant model seems to be avail-

3.3. RETRIEVAL, ALIGNMENT, AND CLUSTERING OF MODELS AND DATA SETS

able for the investigation of the arachidonic acid pathway. However, additional models from Yang *et al.* are available in the literature [Yang et al., 2008]. These models are included in the non-curated part of the BioModels Database and can not be found by the retrieval without further computational steps as they are lacking proper annotations.

3.3.4 Matching Systems Biology models

The retrieval of models similar to a given one and their further comparison can be important during various phases of the modelling process. Before model construction is started one can ask, which models describing a particular process are already available. During the refinement of a model one might be interested in the fact whether models exist, which describe a certain pathway using a different model structure or including further reactions. Furthermore, once a model has been established one can look for models complementing a reaction network by containing additional processes which are not part of the query model.

In the following, results from different applications making use of my vector-based similarity measure will be shown. These results directly address the question of how the aforementioned problems can be dealt with.

3.3.4.1 Retrieving MAP kinase cascade models

A first question during the construction of a mathematical model is whether there are similar or overlapping models available in the literature. This question can either be answered by an extensive literature search or by querying a database of curated existing models. As an illuminating example for the capabilities of the developed web tools, I search for models of MAP (Mitogen-activated protein) kinase cascades. For this purpose I select the kinase cascade model described by Huang & Ferrell [Huang and Ferrell, 1996] (BioModel 9) and start a model retrieval from this model. Figure 3.7 shows the results of this retrieval. The first 14 retrieved models (or 18 out of the first 20) are all MAP kinase cascades or include parts of it. Retrieved models which do not include any elements of a MAP kinase cascade share very unspecific annotations with the Huang model. These include Biological Concepts like “protein phosphorylation” (Gene Ontology term GO:0006468) or “phosphoprotein phosphatase” (EC number 3.1.3.16). As the Huang model contains some of these general annotations, which are relatively common in the BioModels Database, the p-value of the overlap scores is quite high for non-MAP kinase related models.

3.3. RETRIEVAL, ALIGNMENT, AND CLUSTERING OF MODELS AND DATA SETS

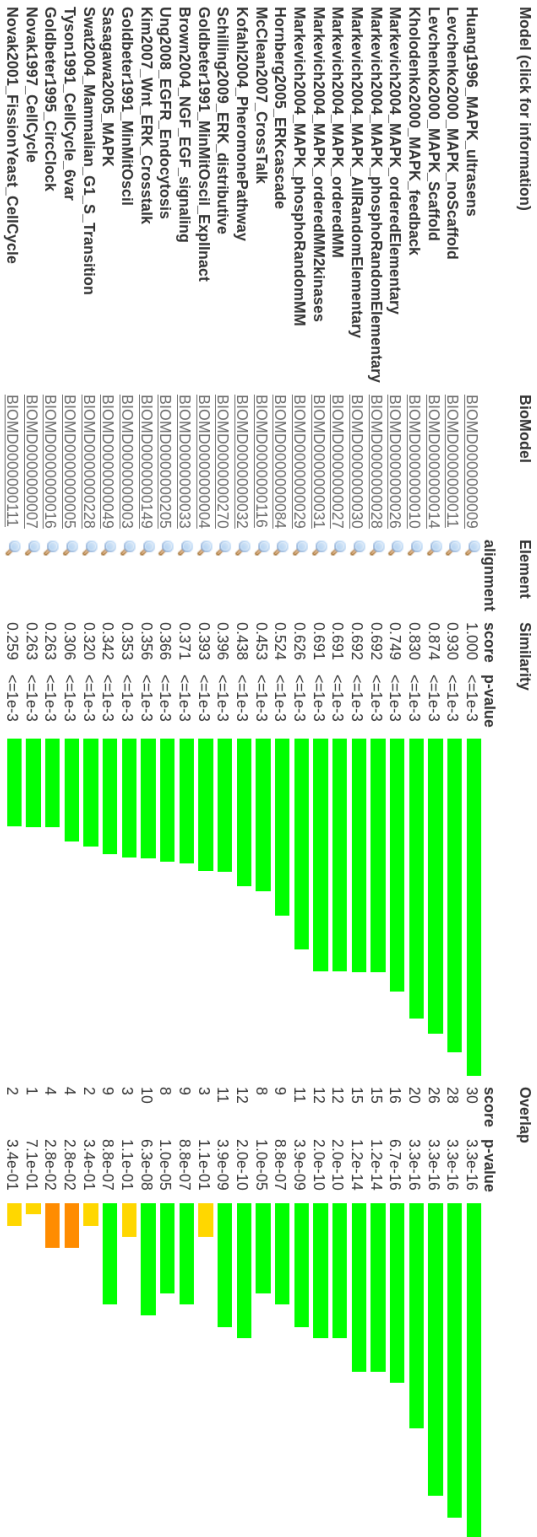


Figure 3.7: BioModels similar to the MAP kinase cascade from the publication of Huang & Ferrell (BioModel 9). The retrieval shows almost exclusively MAP kinase cascades in the first results. When disregarding models showing an insignificant p-value for the overlap score, the results only contain MAP kinase related models.

3.3. RETRIEVAL, ALIGNMENT, AND CLUSTERING OF MODELS AND DATA SETS

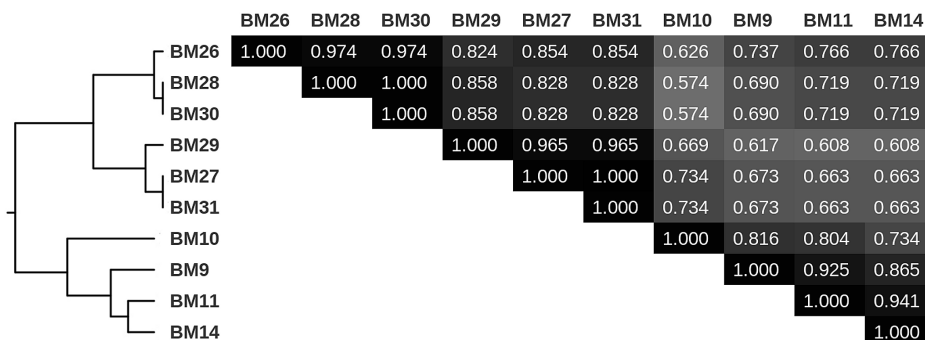


Figure 3.8: Agglomerative clustering using average linkage of the first ten models retrieved in the previous step (see Figure 3.7). The dendrogram on the left visualises in which order the models are clustered together and the matrix on the right is a matrix of pairwise similarities of the models.

The results of this retrieval step show that the similarity as well as the statistical comparison with the null model works in practice. Since both measures show different results (the similarity being a measure of whether the models have the same content and the p-value of the overlap being a measure of how relevant the finding of common Biological Concepts shared by both models is) both measures are kept on the model retrieval page of our web application (<http://semanticsbml.org>).

3.3.4.2 Clustering of MAP kinase cascade models

The second question I try to answer with our tool is how the retrieved models relate to each other. Some of them might differ in the Biological Concepts they describe and others might differ in their degree of detail. To assess the question which models are most related in a set of retrieved models, one can perform a clustering using the vector-based similarity measure.

As an example I cluster the first ten models from the retrieval step (compare Figure 3.7). Figure 3.8 shows the results of an agglomerative clustering using average linkage. The results of this clustering show three important points. First, models from the same publication cluster together (models 26–31 stem from the publication of Markevich *et al.* [Markevich et al., 2004] and models 11 & 14 have been taken from Levchenko *et al.* [Levchenko et al., 2000]). Second, clustering divides the Markevich models into the ones using enzymatic rate laws (27, 29 & 31) and the ones describing the MAP kinase activation by elementary reactions (26, 28 & 30). Yet, the models have a high similarity in the Markevich cluster. This shows that a few annotations, which are specifically used when describing a process using a certain formal-

3.4. DISCUSSION

ism, do not change the high overall similarity of these models. And third, clustering further distinguishes between full MAP kinase cascade models (9, 10, 11 & 14) and the models describing only parts of it.

3.3.4.3 Simple alignments of MAP kinase cascade models

As a final application I use the vector-based similarity measure to align two MAP kinase cascades, the BioModels 9 and 84 [Hornberg et al., 2005b]. These alignments can be used for the detailed comparison of the structural features of models. They can reveal which parts of a pathway are shared between two models or to which degree of detail processes are described by the models.

For the purpose of aligning two models, I assign each model element a feature vector. This vector is built only from the annotations of this element. Then pairwise element similarities between both models are computed using the vector-based similarity measure. Finally, a greedy matching of the elements is performed. Here, pairs having the highest similarities are matched successively as long as they have not been matched already. This process stops, when all elements from one model have been matched or when the similarity drops below a certain threshold.

Results of the alignment of BioModels 9 and 84 are shown in Figure 3.9. Both models contain the activation of three kinases. The Huang model describes these steps using mass action kinetics while the Hornberg model uses enzymatic rate laws. Furthermore, the Hornberg model also contains the activation of the upstream receptor, which is not included in BioModel 9.

Using such an alignment one is able to visualise the commonalities and differences between two models. This can be useful to find structurally different models as an alternative description of a system or to find relevant extensions to a model. Furthermore, this application underlines the versatility of the vector-based similarity measures by the fact that they are extensible enough to be used to compare single model elements instead of complete models.

3.4 Discussion

3.4.1 Retrieval of models and data sets

I have developed a set of similarity measures and have shown various applications for them. These applications can be useful in the construction of computational models that can be used for drug target identification. Out of the shown example uses the retrieval of models describing certain Biological

3.4. DISCUSSION

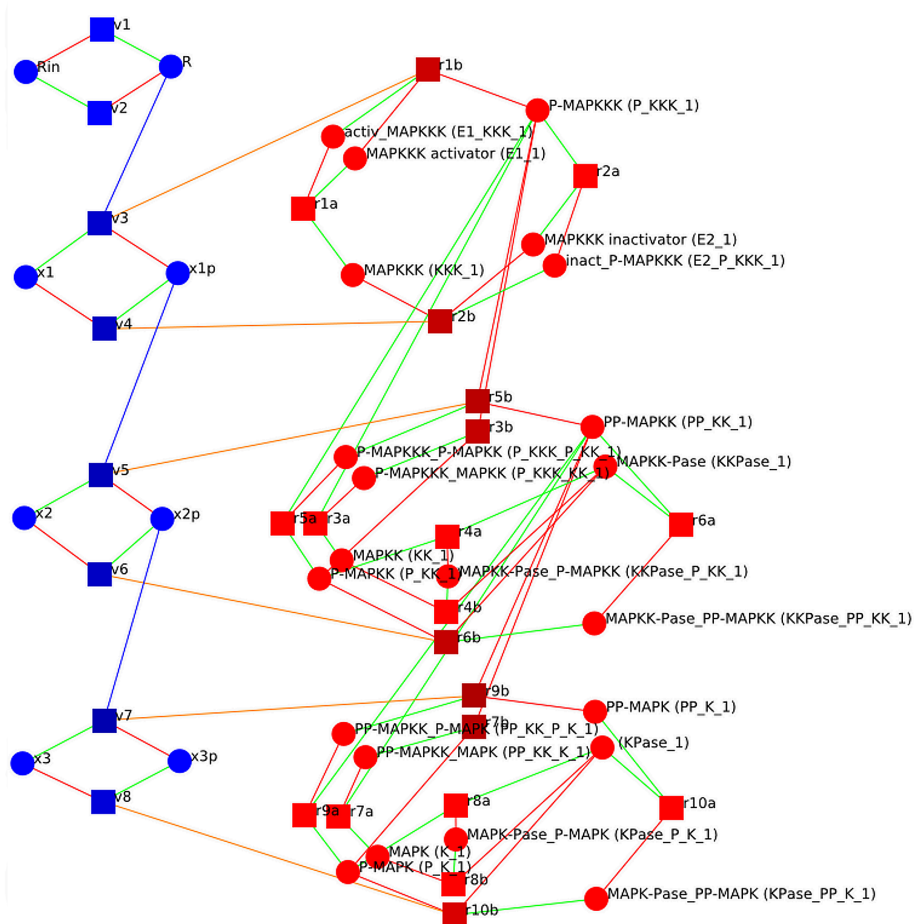


Figure 3.9: Model alignment of the BioModels 9 [Huang et al., 2005] (red) and 84 [Hornberg et al., 2005b] (blue). Circles denote compounds in the network while squares show the reactions converting them. The elements of a reaction network are connected by substrate (green), product (red), and modifier edges (blue). Orange edges connect elements from two models that have been matched.

3.4. DISCUSSION

Concepts is the most important application. Retrieved models can provide a good starting point to all modelling efforts, they can be used to find alternative descriptions of processes possibly leading to an altered behaviour, and they might serve as model extensions describing additional processes. As shown in the results section of this chapter, the vector-based similarity measure can be used to compare models and data sets. Therefore, also the retrieval of relevant data for a given model can in principle be automatised. The retrieved data can then be used to estimate unknown parameters in a model or to further refine it.

In conjunction with the standardisation of models the tools discussed in this chapter open up the possibility of an easy model reuse. As Systems Biology models will further accumulate in the future, there will be a point in time at which it is not possible to manually keep an overview on existing models describing a certain pathway. Thus, computational methods have to be applied to keep the knowledge stored in these models accessible.

3.4.2 Criteria for model similarity

An important question during the development of a similarity measure is the selection of the criteria contributing to this similarity. In the measures introduced in this chapter, the similarity is only determined by the biological content of the model. Therefore, my measures address the point whether compared models describe the same things rather than whether they describe them in the same way. The preference-based measure further uses information on how semantic annotations are distributed in the model, but this is not considered in the vector-based measure.

My measures completely neglect any information on the structure of a model and the particular mathematical formalism used in it. This behaviour is intended as it allows the user to discover different formulations of the same biological process as for example compared in the publication of Markevich *et al.* [Markevich et al., 2004]. Even though the mathematical content or the way in which a model describes a certain pathway is not considered explicitly, a model's formalism can contribute to the way annotations are used in a model, thereby affecting the similarities calculated using my measures. However, information on a model's formalism could be easily included into my similarity measure by using the semantic information stored in SBO (Systems Biology Ontology) terms. These terms can be used in SBML models to indicate the mathematical meaning of variables and formulas and they could be included into the similarity measures in the same ways as other Biological Concepts.

3.4.3 Quality of the presented measures

I have shown that the vector-based measure works in a number of different applications which are relevant to the modelling community. The representation of models or data sets as vectors opens up a whole range of possible new applications like biclustering, principal component analysis [Pearson, 1901], or independent component analysis [Comon, 1994]. Furthermore, the easy structure of the vector-based measure makes it versatilely applicable. It can be used to compare SBML models as well as computational models in other formats, experimental data sets, or plain lists of annotations.

In order to assess the significance of the similarities gained from this measure, I have developed a simple null model and shown ways how to calculate p-values for different measures. This null model is able to detect whether Biological Concepts occur together in a model “on purpose” rather than by chance. Therefore, this p-value can detect whether modellers have tried to describe the same processes or pathways.

The preference-based measures further take into account how annotations are distributed in a model (which model elements are annotated with which BCs). Because this information did not seem to be important in my large scale measure comparison and because the preference-based measures can only be used to compare computational models, I did not include them into our web tools. However, they are mentioned in this work as they might become important as soon as more models become available and a bigger emphasis has to be put on the model’s structure.

Details on the similarity measures for BCs did not play a big role in the large scale comparison. This might stem from the fact that the BioModels have been annotated by the same people, which preferably reuse the same annotations. When models from different sources are compared, details of the BC similarity measure will probably become more important. Nevertheless, the results in Table 3.4 show small positive changes in the quality of the measures when information from the libSBAnnotation and when the distance and the depth factor are taken into account. Including the information content of BCs into the similarity measures did not improve their quality, which is in agreement with the results of Li *et al.* [Li et al., 2003]. Knowledge on term frequencies is, however, not ignored. It still contributes to the p-value of the overlap score and affects which retrieved models should be regarded as significantly similar.

3.4. DISCUSSION

3.4.4 Limitations of the current method

A valid criticism of my similarity measures is that they do not take structural information into account. Even though a purely annotation based approach works in the applications I have shown here, it might not be suitable anymore once more models are available. Future applications will probably have to take structural aspects of the models into account to further refine the retrieval.

This problem also becomes apparent during the alignment of models. The greedy matching which I have applied in conjunction with the vector-based similarity measure will in principle randomly choose to align model elements carrying the same annotations. An example case in which this alignment does not work is when models containing proteins in different phosphorylation states are aligned. Here the annotations only give information on the identity of the protein and the fact that it has been phosphorylated. In order to distinguish between different protein species in distinct phosphorylation states, structural information of the reaction network can be taken into account. Model merging becomes important in the process of constructing large mathematical models that have more power in predicting good drug targets in a network. As a proper matching of model elements is a prerequisite in model merging, I will improve the element similarity in the next chapter by considering structural information.

3.4.5 Conclusion

The similarity measures developed in this chapter work well in practice for the applications I have shown here. Once the amount of information stored in the form of computational models increases beyond the point where a human modeller can have a complete overview on his field, our tools for automated retrieval, comparison, and alignment will play a major role for Systems Biologists. This approach might even become equivalently useful to System Biology as tools like BLAST [Altschul et al., 1990] became to scientists dealing with sequence data. Especially in the process of model driven drug target identification where all publicly available data should be integrated in order to make the most appropriate and biologically relevant predictions our tools can provide a significant contribution to the way modellers work.

Detailed acknowledgements

I would like to thank Falko Krause for initiating work on this project and for doing parts of the programming on the <http://semanticsbml.org> web-

3.4. DISCUSSION

site. Furthermore, I would like to thank Wolfram Liebermeister for general supervision, for proposing the Bayesian estimation of the p-values, and for the idea on how the normalisation for the vector length could be included into the p-value calculation of the overlap score.

3.4. DISCUSSION

Chapter 4

Incorporating structural information into semantic similarity measures

Contents

| | | |
|------------|----------------------------------------------------------------------------|------------|
| 4.1 | Aligning biochemical networks | 100 |
| 4.1.1 | Comparing biochemical networks | 100 |
| 4.1.2 | General differences in network comparison algorithms | 101 |
| 4.1.3 | Merging Systems Biology models | 102 |
| 4.2 | Distributing semantic information in biochemical networks | 103 |
| 4.2.1 | Feature propagation in biochemical networks . . . | 105 |
| 4.2.2 | Semantic propagation for merging network models | 108 |
| 4.2.3 | Predicting missing annotations in biochemical networks | 109 |
| 4.2.4 | Implementation of propagation methods for SBML models | 109 |
| 4.3 | Applications of improved model alignments . . | 111 |
| 4.3.1 | Improvements in the alignment of MAPK models . | 111 |
| 4.3.2 | Randomised removal of annotations and large scale analysis | 113 |
| 4.3.3 | Predicting annotations in a glycolysis model . . . | 115 |
| 4.3.4 | Merging arachidonic acid pathways | 117 |

4.1. ALIGNING BIOCHEMICAL NETWORKS

| | |
|----------------------------------------------------------|------------|
| 4.4 Discussion | 118 |
| 4.4.1 Combining structural and semantic information . . | 118 |
| 4.4.2 Assessing the quality of the proposed measures . . | 119 |
| 4.4.3 Comparison to existing approaches | 120 |
| 4.4.4 Conclusion | 121 |

4.1 Aligning biochemical networks

4.1.1 Comparing biochemical networks

The similarity measures for models and model elements, which have been discussed in the previous chapter, ignore any kind of direct structural information contained in the model. This behaviour makes the measures applicable more broadly but it can also lead to a lack of specificity in the model retrieval and to incorrect model alignments under certain circumstances. In order to compensate for this problem of my similarity measures I will provide extensions to them which directly incorporate a model's topology.

Different graph theoretical and heuristic approaches have already been applied in the field of computational biology to compare different types of networks. These comparisons fall into three different categories [Sharan and Ideker, 2006]: network alignment, integration, and querying. Network alignment is a process in which two networks of similar size are globally compared in order to identify similarities and differences. Network integration combines networks, which can contain different types of information, to detect new or support existing information. Network querying tries to find approximate occurrences of a small motif in a large network. These different approaches have been used in a number of applications.

Network alignment has been used to identify protein-protein interactions (PPI) [Matthews et al., 2001] or regulatory interactions [Yu et al., 2004] conserved across species. The alignment of multiple PPI networks has led to new information about protein functions and protein interactions [Sharan et al., 2005]. Furthermore, network alignment has been used to identify genes which are in close proximity on the genome and catalyse reactions involved in the same pathway [Ogata et al., 2000].

Network integration has been applied to infer PPIs by integrating interactome data, protein domain data, expression data, and functional annotations [Rhodes et al., 2005] or to infer enriched interaction motifs from

4.1. ALIGNING BIOCHEMICAL NETWORKS

PPIs, genetic interactions, transcriptional regulation, sequence information, and gene expression data [Zhang et al., 2005].

Network querying has successfully been used to query genome scale metabolic networks with metabolic pathways in order to find subgraphs with a similar ordering of enzymes [Pinter et al., 2005].

4.1.2 General differences in network comparison algorithms

While the applications of network integration usually follow diverse computational approaches depending on the kind of data the integration should result in, applications of network alignment and querying are more comparable. Even though algorithmic details can be quite different [Li et al., 2007, Singh et al., 2007a], all alignment and querying algorithms consist of two distinct steps: the identification of equivalent nodes, which can be based on diverse information, and the actual alignment, which can make use of different algorithms requiring the compared networks to be of a certain topology.

Depending on the application, the similarity of single nodes in the network needs to be more or less elaborate. In cases in which one does not have detailed information on the nodes (e.g. in the comparison of protein sequences with the nodes being single amino acids) just their labels are compared [Needleman and Wunsch, 1970]. A similarity measure on the labels might be as simple as $\sigma = 1$ for equal labels and $\sigma = 0$ otherwise or a complete pairwise similarity matrix can be used [Dayhoff and Schwartz, 1978]. For cases in which the number of labels is relatively small, such similarity matrices can be constructed. If the number of labels grows further, one has to rely on other information to automatically compute similarities. One possible information source are semantic annotations which can be related to each other through their biological meaning, e.g. “Enzyme Classification” numbers for comparing enzyme functions [Tohsato et al., 2000]. Other potential information sources are structural features of the compared molecules [Hattori et al., 2003] or protein sequence similarity [Kelley et al., 2003].

Depending on the structure of the compared graphs the alignment problem has a different complexity. Early algorithms started with the alignment of simple paths and used ideas from the alignment of nucleotide or protein sequences [Kelley et al., 2003, 2004, Shlomi et al., 2006]. This has been extended to trees [Pinter et al., 2005] (using the approximate labelled subtree homeomorphism algorithm [Pinter et al., 2004]) and general graph structures [Yang and Sze, 2007].

4.1. ALIGNING BIOCHEMICAL NETWORKS

More details on the algorithms used in different network querying algorithms can be found in a recent review [Fionda and Palopoli, 2011].

4.1.3 Merging Systems Biology models

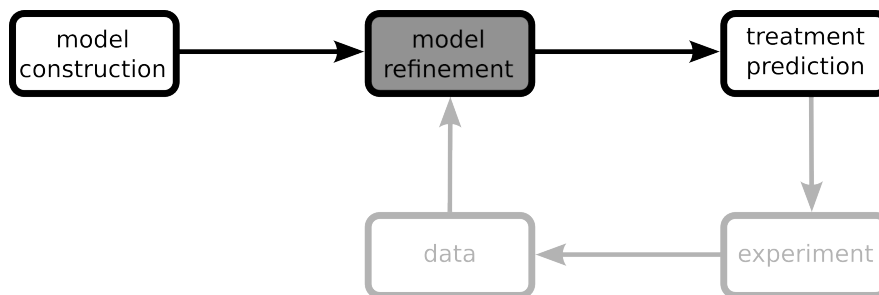


Figure 4.1: Current position in the workflow of applying Systems Biology methods to pharma research and development.

The integration of new information into a mathematical model is a pivotal step in the cycle of Systems Biology as shown in Figure 4.1. An example of this information integration is the merging of models describing distinct, relevant processes. The merging of Systems Biology models for the purpose of producing more comprehensive models has already been discussed in the literature. Comparable to the aforementioned approaches, these merging heuristics can also be divided into consecutive steps: the identification of similar model elements, the matching of elements based on their similarity, and the resolution of possible conflicts stemming from model combination.

KEGGConverter [Moutselos et al., 2009] is a tool which merges pathways from KEGG [Kanehisa et al., 2008] and produces SBML models from them. Since these models stem from the same resource, the identification of identical compounds and reactions can be established via their IDs. Furthermore, the single pathways do not contradict each other and the merging is straightforward.

Goodfellow *et al.* [Goodfellow et al., 2010] have developed a tool which merges SBML models on the basis of their XML code. The way how similar elements across the models are discovered depends on the specific type of element that is compared but is usually quite simple, e.g. in the case of *species* the ID and the name attribute have to match. In the end, a heuristic is applied to fix some of the most common problems arising from the merging of SBML models.

The approach of Randhawa *et al.* [Randhawa et al., 2009, 2010] subdivides the idea of model merging into distinct tasks with a different outcome:

4.2. DISTRIBUTING SEMANTIC INFORMATION IN BIOCHEMICAL NETWORKS

composition is a process in which elements from models are linked but the individual models are kept intact, fusion combines models irreversibly into a single new model, and flattening turns a composed into a fused model. Depending on which of the processes is used to merge models, the process varies. However, the identity of model elements is in each case established via their names or predefined by the user.

PInT [Wang et al., 2010b] is another tool to merge SBML models. Here, the model elements are compared by their annotations. If these are completely identical, the elements are merged. The final processing of the merged models is also in this case a heuristic addressing a set of potential conflicts stemming from the merging process.

The software semanticSBML [Schulz et al., 2006, Krause et al., 2010] is similar to PInT but uses an element similarity measure [Schulz et al., 2011] to judge whether two model elements are identical. This approximation simplifies the merging process when two models from different sources, which have not been annotated by the same people, are combined.

4.2 Distributing semantic information in biochemical networks

The aforementioned methods for network alignment and querying have been developed mostly for the analysis of PPI networks. When these methods are employed to compare Systems Biology models, they are facing a slightly different challenge. The idea behind them is to find subgraphs of similar structure in the query and the large network, which are equivalent to each other (isomorphic) after a small number of edit operations (node insertions and deletions). Depending on the number of allowed operations the computational costs of the algorithms can become quite large. Therefore, this number is usually kept small.

When kinetic models of reaction networks are compared by a modeller diverse structures can be regarded as equivalent. Figure 4.2 shows such equivalent structures, which require several edit operations to become isomorphic. Already for small networks, the number of necessary operations exceeds the limit imposed by computational feasibility, which heavily reduces the applicability of the methods developed for PPI networks. One way to circumvent this restriction is to especially allow for those edit operations interconverting structures from Figure 4.2, e.g. replacing enzymatic reactions by their elementary reaction steps, thus keeping the number of needed operations minimal [Gay et al., 2010]. A second idea, the inclusion of structural

4.2. DISTRIBUTING SEMANTIC INFORMATION IN BIOCHEMICAL NETWORKS

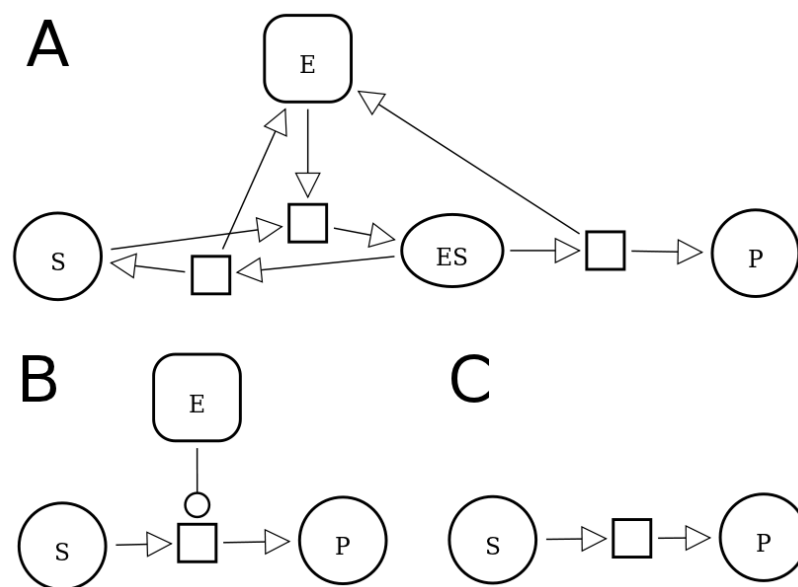


Figure 4.2: Structures of reaction networks regarded as similar in Systems Biology models. The three networks describe a single irreversible enzymatic reaction using (A) mass action kinetics or (B) Michaelis-Menten kinetics with or (C) without explicitly modelling the enzyme. This Figure is based on Figure 1 from [Gay et al., 2010].)

4.2. DISTRIBUTING SEMANTIC INFORMATION IN BIOCHEMICAL NETWORKS

information into the similarity of nodes will be discussed here.

4.2.1 Feature propagation in biochemical networks

4.2.1.1 General idea behind feature propagation

The idea of feature propagation is based on the following observation. If an interaction exists between two proteins, there is an increased probability of both proteins sharing the same functional annotation [Schwikowski et al., 2000]. Thus, information on interacting proteins can be used to annotate proteins with unknown function. In principle, this idea can also be transferred to reaction networks. This is obvious for models of signalling cascades, in which proteins interact, but it is also possible for metabolic networks. Here, the interactions are the information which compound participates in which reaction. Using the transfer of annotations along the interactions, one could annotate reactions with semantic information from their reactants, products, and the enzyme catalysing them.

The concept of propagating semantic information along interactions has already been applied by different tools. FunctionalFlow [Nabieva et al., 2005] proposes a stepwise distribution of annotations along a protein interaction network. Using this distributed annotations, novel protein functions can be predicted. Another example of semantic propagation is discussed in [Singh et al., 2008]. Here, the authors formulate an eigenvalue problem enforcing global similarities between two nodes to follow a relation similar to the similarity propagation formula. More details on predicting annotations in protein-protein interaction networks can be found in [Sharan et al., 2007].

4.2.1.2 Mathematical formulation of feature propagation

I have implemented the distribution of semantic information, which is represented by feature vectors, on the network graphs by a non-mass-conserving diffusion-like process as shown in Figure 4.3 [Schulz et al., 2012]. In this process one defines the change over time of the new, *inferred* feature vector $w_{*,i}$, which denotes the distribution of a certain information i over all nodes of the network, by the formula

$$\frac{d}{dt}w_{*,i} = v_{*,i} + \lambda R w_{*,i} - w_{*,i}. \quad (4.1)$$

This formula contains a production term ($v_{*,i}$) producing semantic information on those nodes, which have been assigned semantic information on the Biological Concept i , a “diffusion” term incorporating the network topology

4.2. DISTRIBUTING SEMANTIC INFORMATION IN BIOCHEMICAL NETWORKS

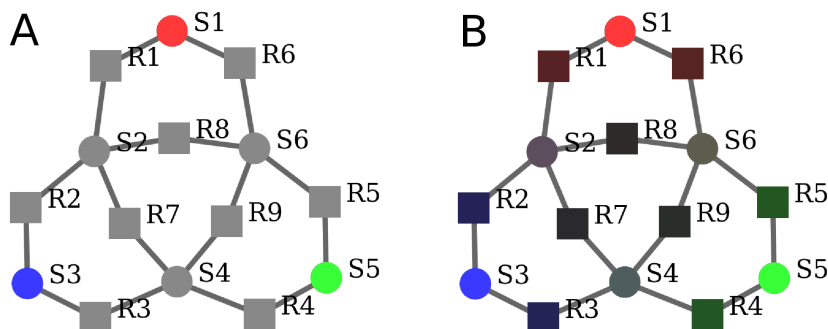


Figure 4.3: Propagation of colour information in a reaction network (circles: compounds, squares: reactions). (A) A network with sparse semantic information (shown as colours red, blue, and green) on three out of its fifteen nodes. (B) After propagation of the colour information in the network every node is assigned a distinct colour determining its identity.

(λR), and a linear degradation term. In this scheme, the matrix R is an $|M| \times |M|$ matrix, where $|M|$ is the number of elements in model M , and λ is a scaling factor, that is determined later. R consists of the propagation weights $R_{ab} = \rho_{ab}^M$ that assume non-zero values $\rho_{ab}^M = \alpha$, $\rho_{ba}^M = \beta$ in cases in which the elements a and b are directly related to each other (e.g. species a participating in reaction b) and $\rho_{ab}^M = 0$ otherwise. The distribution of the inferred features on the network is eventually given by the steady state of the diffusion process ($w_{*,i} | \frac{d}{dt} w_{*,i} = 0$).

Alternatively, the diffusion process can be formulated in terms of vectors $w_{a,*}$ describing all information associated with a certain model element a . This vector can be defined by the implicit formula

$$w_a = v_a + \lambda \sum_{b \in M} \rho_{ab}^M w_b, \quad (4.2)$$

which shows how the inferred features on related elements b contribute to the inferred features on an element a .

4.2.1.3 Observations on feature propagation

In order to get an explicit formula for the computation of $w_{*,i}$, one solves Equation 4.1 for its steady state

$$\begin{aligned} 0 &= v_{*,i} + \lambda R w_{*,i} - w_{*,i} \Leftrightarrow \\ w_{*,i} &= (I - \lambda R)^{-1} v_{*,i} \end{aligned}$$

4.2. DISTRIBUTING SEMANTIC INFORMATION IN BIOCHEMICAL NETWORKS

If λR has no eigenvalues whose absolute value is larger than 1, the matrix inverse can be replaced by

$$w_{*,i} = \sum_{k=0}^{\infty} (\lambda R)^k v_{*,i}. \quad (4.3)$$

This can be ensured by setting $\lambda = \frac{1}{2|r|}$, where r is the eigenvalue of R with the largest absolute value.

In the infinite sum each summand describes how much semantic information is propagated from one element to the elements being k relations away. Since λR has no eigenvalue whose absolute value is bigger than one, higher exponentiations of this matrix will have smaller and smaller eigenvalues. Therefore, features are propagated less strongly to those nodes in the network that are more distant.

In case the reaction graph is acyclic and features are only propagated in one direction (e.g. compartments to species and species to reactions) higher powers of λR will become zero and the infinite series in Equation 4.3 will become finite. Furthermore, if ρ values are non-negative and the direct feature vectors are non-negative, too, the values in the propagated feature vectors are non-negative.

Finally, the vectors $v_{*,i}$ and $w_{*,i}$ can be combined in matrices V and W leading to a single equation for the propagated features

$$W = (I - \lambda R)^{-1}V. \quad (4.4)$$

Similar to Equation 3.8, a vector based similarity measure for the inferred feature vectors can be defined by

$$\psi_{mn}^{\text{fp}} = \frac{w_m^{\text{T}} S w_n}{\sqrt{w_m^{\text{T}} S w_m} \sqrt{w_n^{\text{T}} S w_n}}.$$

Propagating pairwise similarities A slightly different idea on how to distribute semantic information in a network is the concept of similarity propagation. In contrast to the feature propagation, which distributes features in the individual models, pairwise similarities between elements from different models are propagated.

Given *direct* similarities between elements in two models M and N , which do not necessarily have to be based on vectors, the *inferred* similarities are defined in analogy to Equation 4.2 by the formula

$$\psi_{ap}^{\text{sp}} = \sigma_{ap} + \lambda \sum_{b \in M, q \in N} \rho_{ab}^{\text{M}} \psi_{bq}^{\text{sp}} \rho_{pq}^{\text{N}}.$$

4.2. DISTRIBUTING SEMANTIC INFORMATION IN BIOCHEMICAL NETWORKS

In this formula a potential similarity of the elements b and q contributes to the similarity of elements a and p . After writing the pairwise similarities into a single vector (as indicated by brackets on the index), the formula for the computation of the propagated similarities can be stated similarly to Equation 4.4:

$$\begin{aligned}\psi_{(ap)}^{\text{SP}} &= \sigma_{(ap)} + \lambda \sum_{(bq)} Q_{(ap)(bq)} \psi_{(bq)}^{\text{SP}} \Rightarrow \\ \psi^{\text{SP}} &= (I - \lambda Q)^{-1} \sigma,\end{aligned}$$

where $Q_{(ap)(bq)} = \rho_{ab}^{\text{M}} \rho_{pq}^{\text{N}}$.

As mentioned above, by an appropriate choice of λ one can ensure that the propagated similarities are non-negative. However, the propagated similarities may become bigger than one. If this behaviour is not desired, the propagated similarities have to be normalised, e.g. by

$$\overline{\psi}_{ap}^{\text{SP}} = \frac{\psi_{ap}^{\text{SP}}}{\sqrt{\max_{b \in M} \psi_{bp}^{\text{SP}}} \sqrt{\max_{q \in N} \psi_{aq}^{\text{SP}}}}.$$

The two different propagation methods are in fact closely related. Similarity propagation also simulates the outcome of a non-mass-conserving diffusion process. Instead of propagating semantic information on the reaction network of a single model, the semantic information is propagated on a graph whose nodes correspond to pairs of elements from two networks. Nodes in this graph are connected by edges in case a relation between each of the model elements exists in both models. This means that the nodes (ap) and (bq) are connected if a relation between the elements a and b and between p and q exists.

4.2.2 Semantic propagation for merging network models

My prior approach to model merging has been based on a greedy pairing of model elements. This greedy pairing is performed by putting the pairwise similarities into a decreasing order and successively matching those pairs with the highest similarity for which none of the elements has already been matched. As soon as the similarities drop below a certain threshold, the matching is stopped. After the pairing of model elements, the matched pairs are merged and potential conflicts in the new merged models are removed by applying some heuristic rules (see [Schulz et al., 2006] for details).

4.2. DISTRIBUTING SEMANTIC INFORMATION IN BIOCHEMICAL NETWORKS

The computation of the pairwise similarities can incorporate structural information contained in both models by using the similarity calculated from propagated features or by using the propagated similarities. Regardless of the fact, which similarity measure is used in this initial step, all further steps can follow the aforementioned scheme.

4.2.3 Predicting missing annotations in biochemical networks

The propagation of features can also be used to predict annotations for model elements lacking semantic information. The basic idea for this step is to align a sparsely annotated model to a well annotated one and transfer semantic information to the matched, non-annotated elements. In order to work perfectly, this step would require a structurally identical, annotated model. Since such models will not universally be available, I instead use all models in the curated part of the BioModels Database to predict new annotations.

Computationally the prediction works by setting up a database of pairs of annotations and propagated feature vectors. For each element a in each BioModel one creates the propagated feature vectors w_a and for all of the element's annotations i one constructs modified feature vectors lacking the corresponding features $w_a^i = w_a - w_{ai} \cdot e_i$, with e_i being the i^{th} canonical unit vector. The pairs of the vector w_a^i and the feature i are then stored in the database.

When predicting new annotations for model elements, one first propagates features in this model and then one uses the similarity measure from Equation 4.5 to search the database for propagated feature vectors pointing into a similar direction. Finally, the features associated with the most similar vectors are presented to the user as potential annotations for the considered model element.

4.2.4 Implementation of propagation methods for SBML models

Up to this point, I have mainly regarded a model as a reaction network consisting of species and reaction. SBML models, for which these methods have been implemented, have many more element types that can have various relations with each other along which semantic information can be propagated. One example is a model containing the same species in two different compartments. The two species might actually carry the same annotations, but might refer to different compartments that carry annotations themselves.

4.2. DISTRIBUTING SEMANTIC INFORMATION IN BIOCHEMICAL NETWORKS

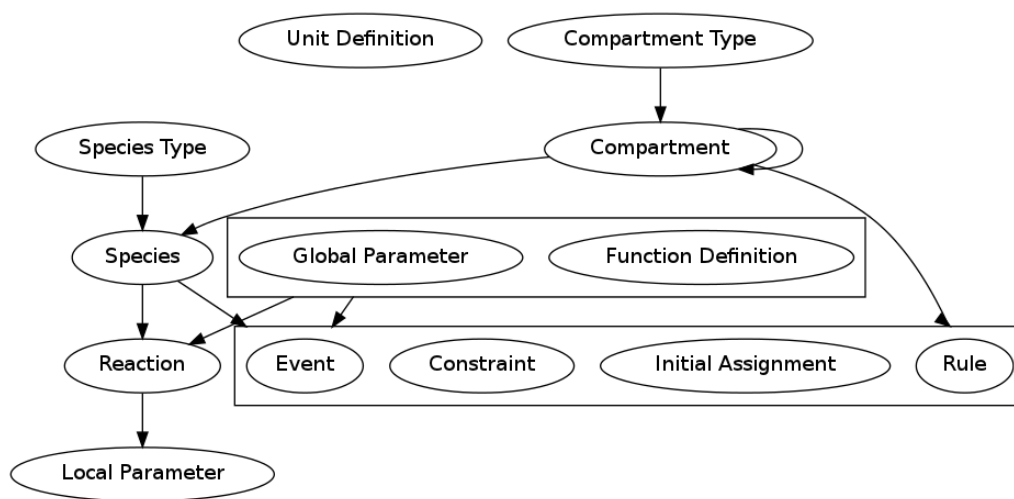


Figure 4.4: Directed structural dependencies between elements of an SBML model. Species are located in certain compartments, which may be sub-compartments of each other, and these can be of a certain compartment type. Along each arrow, semantic information can therefore be transmitted between elements of a different type. The dependencies between compartments themselves do not impose that a potential cycle exists. Compartments may refer to one other compartment as their “outside”, but these references should not involve any circularities.

4.3. APPLICATIONS OF IMPROVED MODEL ALIGNMENTS

Propagating the annotation of the compartment to the species contained in it, gives the two species distinct semantic information.

Figure 4.4 shows a complete overview on the direct relations existing between different types. In principle, information could be propagated along all relations in both directions.

Numerical values So far the choice of the numerical values used for propagating semantic information has not been discussed. For the implementation of my two applications, the alignment of sparsely annotated models and the annotation prediction, I use two different sets of numerical values.

In the alignment routine on <http://semanticsbml.org>, the tool propagates information between all the element types shown in Figure 4.4 in both directions with $\rho_{ab} = \frac{1}{2}$. The fact why these values have been chosen can easily be explained. First, apart from choosing the numerical value of $\frac{1}{2}$ *ad hoc* I have been unable to optimise numerical values due to a lack of multiple reference alignments, which could be used for benchmarking. Second, I propagate along all relations in the network as I would like all semantic information to be (at least partially) available on all nodes in the network.

For the prediction of annotations I have made some alterations to how features are propagated in the network. First, information is only propagated between reactions and their reactants or products. Second, the features are only passed to the next node in the network. This means, that the sum in Equation 4.3 only runs to one instead of to infinity. Using these two modifications increased the performance of the model prediction as it decreases the number of proposed annotations in most examples (data not shown). Furthermore, the decision to propagate information only to the direct neighbours in a network is supported by results for PPI networks, for which the “Markov property” seems to hold, i.e. only the direct neighbours contribute to the identity of a protein [Deng et al., 2003].

4.3 Applications of improved model alignments

4.3.1 Improvements in the alignment of MAPK models

As a first example of how accounting for structural information in a similarity measure can increase the quality of the alignment of two models, I compare the introduced approaches using two MAP kinase cascade pathways. In Figure 4.5, I use the greedy pairing heuristic based on three different similarity measures to align the BioModels 9 and 11 [Huang and Ferrell, 1996,

4.3. APPLICATIONS OF IMPROVED MODEL ALIGNMENTS

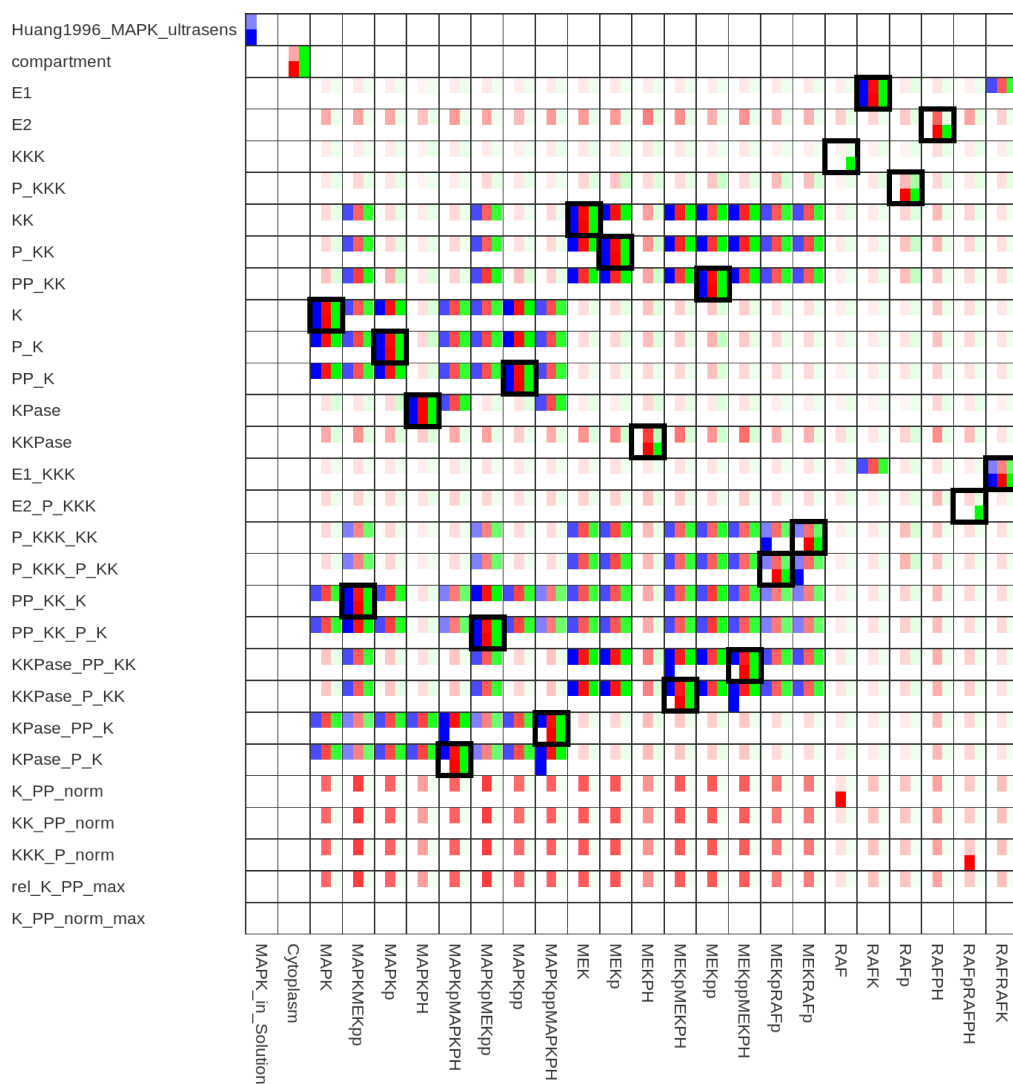


Figure 4.5: Alignment of BioModels 9 (y-axis) and 11 (x-axis) based on similarities produced by different *direct* and *inferred* measures. The six cells per element pair denote the similarity (top row) and the greedy pairing based on this similarity (bottom row) for the direct, vector-based measure (blue), for the measure based on feature propagation (red), and for the propagated similarities (green). Numerical values of the similarities are indicated by colour intensities. Black boxes around element pairs denote a manually curated reference alignment between the species in the two models.

4.3. APPLICATIONS OF IMPROVED MODEL ALIGNMENTS

Levchenko et al., 2000].

A problem of the direct similarity measure with this alignment is that elements representing proteins and protein complexes in different phosphorylation states carry identical annotations, e.g. P_KKK_KK and P_KKK_P_KK in the Huang model and MEKRAFp and MEKpRAFp in the Levchenko model. While elements carrying identical annotations are correctly resolved by both propagation methods, a few differences between both approaches exist. The pairing based on propagated similarities matches the pairs KKK & RAF and E2.P.KKK & RAFpRAFP, which are missed by the two other measures. Nevertheless, the feature propagation performs better on matching the reactions between the two models (matching of reactions is not shown). Here, only feature propagation is able to resolve the correct matching of the reactions with the IDs r7a & Reaction19 and r9a & Reaction25.

In this example both propagation methods seem to improve the quality of the alignment. Nevertheless, both approaches still show some small shortcomings and one cannot state that either of them has an advantage over the other.

4.3.2 Randomised removal of annotations and large scale analysis

In order to further validate the results of the single alignment and to discriminate between both propagation methods, I evaluate the different measures in a larger analysis. For this purpose, I randomly removed elements from BioModel 9 and again tried to align it to the Levchenko model. This has been repeated ten times for different numbers of annotations to be removed and the quality of the resulting predictions is shown in Figure 4.6.

When measuring the quality of the matching in terms of the recall (the number of correct matches that have been identified), the power of the alignment based on direct similarities decays linearly with the number of annotations that have been removed from one of the models. For the propagated features the behaviour is similar except from the fact that the linear decay starts from a higher recall for few missing annotations. In the case of propagated similarities the decay in the recall seems to be biphasic. First, up to 30 removed annotations the decay is linear but less steep than in the case of the propagated features. Then, the recall decays hyperbolically for higher numbers of removed annotations and even for a single remaining annotated element one fifth of the model can be aligned correctly.

When the quality of the matching is measured in terms of the precision (the number of predicted pairs being correctly matched) the alignment based

4.3. APPLICATIONS OF IMPROVED MODEL ALIGNMENTS

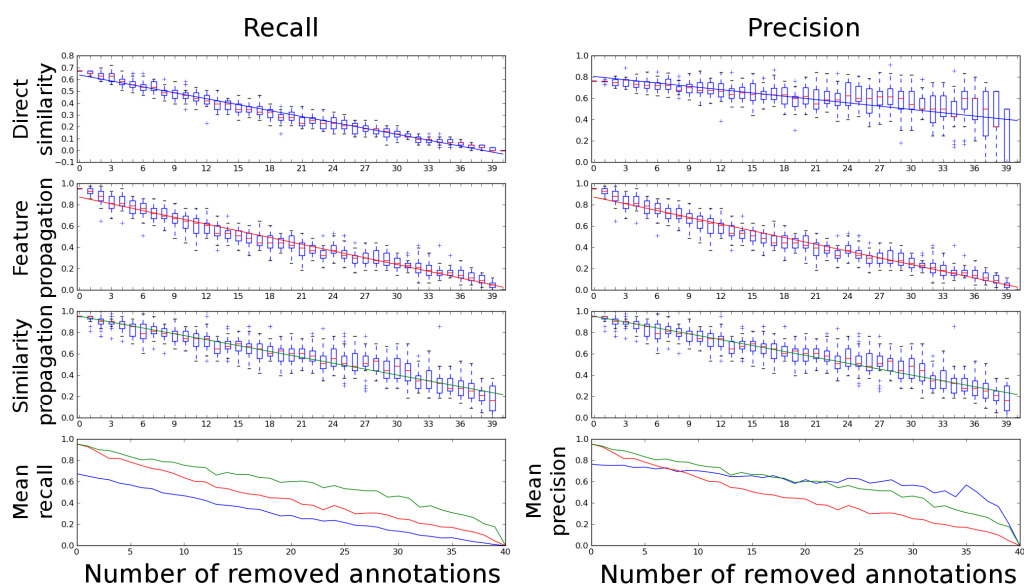


Figure 4.6: Quality of the alignment of BioModels 9 and 11 after removal of random annotations from the Huang model. Depicted are precision and recall of the proposed element pairs (compared to the reference alignment from Figure 4.5). For each number of elements cleared from its annotations ten repetitions have been performed. Regression lines have been added to the first three plot lines to visualise the mostly linear decline in precision and recall. Furthermore, a fourth line of plots visualises the mean values of precision and recall for all three methods.

4.3. APPLICATIONS OF IMPROVED MODEL ALIGNMENTS

on direct similarities scores already quite well. Here, no predictions about elements lacking annotations are made and therefore only the ambiguity in the matching of elements carrying identical annotations influences the precision negatively. For the propagated features the decay in precision seems to be linear along the full range of removed annotations. But even though the precision is higher for few removed annotations than for the matching based on direct similarities, its quality drops faster. In the case of propagated similarities, the behaviour is again biphasic as for the recall. It drops linearly for up to 30 removed annotations but less steep than the precision of the matching based on feature propagation. For higher numbers of removed annotations this behaviour is again hyperbolic and even in the cases where a single annotation is left in one of the models one fifth of the predicted matchings are correct. Generally, even though both propagation methods perform better in terms of the recall, they perform worse than the direct measure in terms of precision. This can be easily explained by the fact that they also try to match non-annotated elements, which they might do incorrectly. Therefore, their predicted matchings will contain more errors than the direct matchings, especially if one of the matched models is annotated sparsely.

The same kind of analysis has been performed for other model pairs. Since the results do not change qualitatively, they are not shown here. It would in general be interesting to run this kind of analysis on a larger scale with lots of different model pairs. Unfortunately, a manually curated matching of the elements is required in order to measure the performance of the different approaches. As these matchings are currently not available such a kind of analysis has to be postponed.

4.3.3 Predicting annotations in a glycolysis model

In order to test the quality of the annotation prediction I perform a similar kind of analysis. I repeatedly remove a varying number of annotations from the glycolysis model of Hynne [Hynne et al., 2001] and try to repredict the removed annotations. Instead of a signalling cascade model as in the previous examples, I use a model describing a metabolic network. In contrast to the aforementioned models, this model does not contain multiple species with identical annotations.

For each element whose annotations have been randomly removed I use its propagated feature vector to search for relevant annotations in the feature-vector database. Given the topmost hits (i.e. the features associated with vectors pointing into a similar direction), I evaluate the position of the correct annotation in the retrieved list.

The results of this evaluation for BioModel 61 are given in Figure 4.7.

4.3. APPLICATIONS OF IMPROVED MODEL ALIGNMENTS

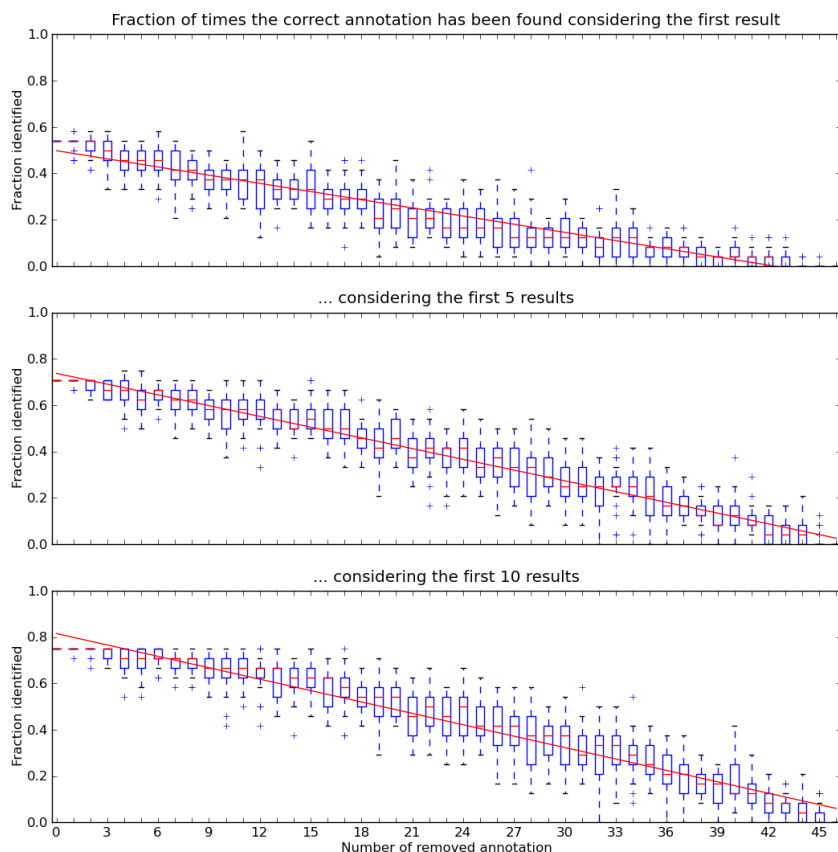


Figure 4.7: Large scale analysis of the quality of the annotation prediction. Certain numbers of randomly chosen annotations (x-axis) have been removed from BioModel 61 [Hynne et al., 2001], a model of glycolysis, in ten repetitions. After the removal, I try to predict the missing annotations and compute the fraction of times (y-axis) I have been able to find the annotation in the first n results (different plots visualising $n = 1, 5,$ and 10).

4.3. APPLICATIONS OF IMPROVED MODEL ALIGNMENTS

These results show two points: First, the quality of the prediction seems to be linearly dependent on the amount of annotations which are already present in the model. The more annotations are available, the better the prediction of novel annotations works. Also the quality of the model matching based on feature propagation has shown a linear dependence on the number of annotations present in the models. Although this behaviour seems to be recurring it cannot be easily explained because of the complexity of the feature propagation and the similarity measure. Second, the correct annotation is in most cases contained in the first ten suggested annotations. The plot in Figure 4.7, which depicts the fraction of correct annotations being in the first ten suggested ones, did not differ significantly from a curve for the first 50 suggestions. Furthermore, about half of the times that the correct annotation is included in the first ten retrieved ones, it is the first. While this underlines the quality of my suggestion heuristic, the method is insufficient to predict annotations completely automatically. Nevertheless, the method is adequate to suggest annotations in a user interface, as it is implemented on <http://semanticsbml.org>.

4.3.4 Merging arachidonic acid pathways

As one of the running examples throughout this thesis, a comprehensive model of the arachidonic acid pathway is constructed and analysed. For this purpose I perform a retrieval step on the non-curated part of BioModels Database starting from an initial annotated model that had previously been identified [Yang et al., 2007]. As the models in the non-curated part have not been assigned semantic information, they have to be annotated automatically beforehand (details omitted). This retrieval step leads to the identification of three further models describing the arachidonic acid pathway. In contrast to the query model, which describes the pathway in granulocytes, the more recent models describe it in granulocytes as well as in endothelial cells and platelets [Yang et al., 2008]. As the retrieval models are an improvement on the first model, only these are used and combined into a single model.

My objective behind running the drug target identification in the three different cell types, is to find a potential treatment, which can inhibit the production of pain mediators in all types simultaneously. Thus, when the three pathways are combined into a single model, I do not couple the pathways' differential equations. After combining the models one ends up with a model containing three arachidonic acid pathways in three different compartments. In further steps the automatically assigned annotations have been manually curated to ensure that model elements are described by appropriate semantic information. Additionally to the models published in BioModels Database,

4.4. DISCUSSION

the publication of Yang *et al.* contains four further parameter sets which fit their experimental equally well. Making use of these parameter sets one ends up with five different models, each describing the arachidonic acid pathway in three distinct cell types.

4.4 Discussion

4.4.1 Combining structural and semantic information

What have I achieved so far? The models I am working with need to carry specific semantic information which explain the Biological Concept behind each model element. Without this kind of information, many of the methods introduced in the scope of this thesis would not work. Nevertheless, assigning annotations to all elements in a model can become an excessive amount of work. This amount of work increases even further when the assigned semantic information is critical for the success of later methods and, therefore, has to be as precise as possible.

Because of the additional effort that has to be put in the annotation procedure, many available models do not contain semantic information. There are two ways in which this problem can be handled: First, a tool supporting the automated assignment of semantic information to elements in a model could be developed. Second, methods in need of completely annotated models, e.g. alignment algorithms, could be extended by working on partially annotated ones. The approaches introduced in this chapter address both of these ideas using a similar mathematical method.

Propagating information The propagation of semantic information in a sparsely annotated biochemical network allows us to assign unique feature vector to most of the elements in a model. These feature vectors describe the distribution of semantic information within the network. For the purpose of propagating the information in the network a non-mass-conserving diffusion-like process is applied to the features, which, judging from the positive results, seems to combine structural and semantic information appropriately.

This approach can further be extended to propagate information about the similarities of related pairs of model elements. Using this kind of propagation in the alignment of different models leads to improved results, but its computational demands are by far higher ($\mathcal{O}(|M|^3 \cdot |N|^3)$ instead of $\mathcal{O}(|M|^3 + |N|^3)$, where $|M|$ and $|N|$ denote the numbers of elements in the two models). However, similarity propagation is more broadly applicable

than feature propagation because it can also be applied to similarity measures which are not vector-based.

Apart from its use during the process of merging sparsely annotated models, feature propagation can also be of help to assign semantic information to model elements. This ability has been added to our web tool at <http://semanticsbml.org>, which does not only allow to add MIRIAM-compliant annotations to an SBML model but also allows for query-based retrieval of annotations by their names. The inclusion of the annotation prediction into this interface adds another layer of comfort to the annotation process and is able to reduce the effort associated with it.

4.4.2 Assessing the quality of the proposed measures

Quality of model alignment The quality of the model alignment based on the different propagation schemes is higher than for the cases in which the direct pairwise similarity has been used. Although the alignments differ between the propagation schemes, they appear to be stable across a wide range of parameter values for the different relations in the R matrix. One possible explanation for this behaviour is the adaption, which is done through the choice of the λ parameter (half of the inverse of the largest absolute value of an eigenvalue).

Although the results of an alignment might not be very sensitive to the choice of the numerical values in the R matrix, it should be a matter of discussion along which relations in an SBML model semantic information should be propagated. In cases in which all relations propagate information, some elements might be matched without sufficient supporting information. This can happen if two different compartments each contain a species sharing a single annotation. This single annotation contributes positively to the similarity of the species, which in turn propagates to the similarity of the compartments. If such a behaviour is not desired, the threshold for the matching of elements could be increased or, depending on the exact application, selected propagation weights could be altered.

Quality of annotation prediction As shown in the results section of this chapter, the prediction of new semantic information for elements in a sparsely annotated model using my approach provides sensible results. In general, the annotation prediction works best, when all elements in the considered model represent distinct BCs. Therefore, the semantic information in metabolic models is easier to predict than in models of signalling networks. As discussed for the example of BioModel 9, elements can share some or even

4.4. DISCUSSION

all of their annotations even though they describe distinct states of a certain protein. These shared annotations can be a problem for the annotation prediction, because the semantic information which is shared across all these elements might be overrepresented in the internal “feature–vector” database and, therefore, in the results of this method.

4.4.3 Comparison to existing approaches

Applying graph matching methods to Systems Biology models In general, the graph matching algorithms that have been mentioned in the introduction of this chapter, divide the process of the model alignment into two distinct steps: First, pairwise similarities between the models’ elements are computed. These are mainly based on local properties of the nodes such as their labels, their associated sequence information, or their molecular structure. Second, based on these similarities, a matching of the model elements is computed, which respects the individual graph structures. The numerical efficiency of this second step is in most cases dependent on how much the compared graphs are allowed to differ in order to be still regarded as equivalent. In the comparison of Systems Biology models, the differences can be huge. Therefore, either the computation becomes very expensive or a different approach has to be employed to find a suitable model alignment.

Modifications suitable for Systems Biology One possible way out of this dilemma is to relax one of the requirements of the alignment. As soon as one drops the idea that two graphs are supposed to be isomorphic after a certain number of edit operations, this number does not limit the applicability of my method. In order to still incorporate structural information into the computation of pairwise similarities, semantic information is propagated along the network structure. Given the fact that structural information has been considered in the first step of model alignment, the second step – the actual alignment – can be performed in a lazier manner. But apart from the greedy matching heuristic employed in this chapter, the propagated similarities can also be used in combination with the aforementioned graph matching algorithms. By employing the introduced measures, the quality of the alignment of the graph matching methods could be increased, or the measures could be used for filtering interesting regions of a large pathway map for interesting subgraphs.

4.4.4 Conclusion

As demonstrated in the results section of this chapter, feature propagation and similarity propagation do improve the quality of model alignments. Furthermore, feature propagation can be used to predict new annotations. In contrast to methods established for the comparison of PPI networks, the methods introduced in this chapter can also be applied to compare Systems Biology models. They can deal with the problem of aligning models describing processes to a different degree of detail, e.g. models using different kinetics. Furthermore, they are able to distinguish between elements carrying identical annotations.

In conclusion, the similarity measures described in this chapter integrate structural information into the measures from the previous chapter in a consistent manner, thereby resolving aforementioned problems.

Detailed acknowledgements

I would like to thank Falko Krause for his programming work on the web interface of semanticSBML, especially for rewriting the annotation and merging code. Furthermore, I would like to thank Wolfram Liebermeister for general supervision and for proposing the similarity propagation formula.

4.4. DISCUSSION

Part II

Predicting possible drugs and drug targets

Chapter 5

Identifying potential drug targets in ODE models

Contents

| | | |
|------------|---------------------------------------------------------------------------------------------|------------|
| 5.1 | Drug target identification by parameter optimisation | 126 |
| 5.1.1 | Drug target identification is a parameter estimation problem | 127 |
| 5.1.2 | Parameter identifiability | 129 |
| 5.1.3 | Host pathogen systems | 131 |
| 5.2 | Formalising the drug target identification problem | 131 |
| 5.2.1 | Setting up a parameter estimation to solve the drug target identification problem | 132 |
| 5.2.2 | Different optimisation methods and what solutions to expect | 136 |
| 5.2.3 | Parameter identifiability | 138 |
| 5.2.4 | Host-pathogen interactions | 140 |
| 5.3 | Results | 142 |
| 5.3.1 | Implementation | 142 |
| 5.3.2 | Linear pathway | 143 |
| 5.3.3 | Glycolysis in <i>Trypanosoma brucei</i> | 145 |
| 5.3.4 | AA pathway | 148 |
| 5.4 | Discussion | 154 |
| 5.4.1 | What has been achieved in this chapter | 154 |

5.1. DRUG TARGET IDENTIFICATION BY PARAMETER OPTIMISATION

| | | |
|-------|---------------------------------------------|-----|
| 5.4.2 | Comparison to other approaches | 154 |
| 5.4.3 | Biologically relevant results | 158 |
| 5.4.4 | Drug resistance through mutations | 161 |
| 5.4.5 | Outlook | 162 |

5.1 Drug target identification by parameter optimisation

Current state

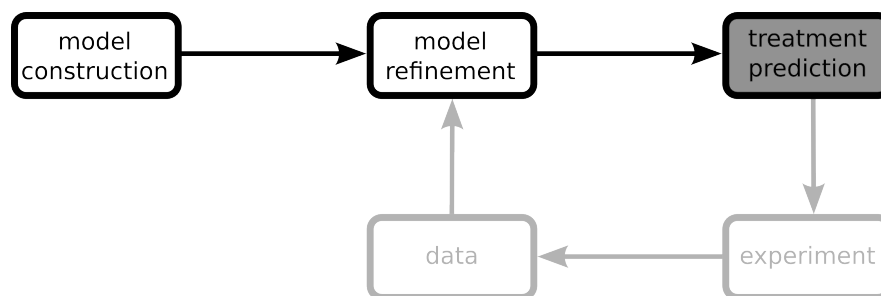


Figure 5.1: Current position in the workflow of applying Systems Biology methods to pharma research and development.

In the previous chapters I have discussed methods for the retrieval of knowledge in the form of computational models and experimental data sets as well as methods for the alignment and integration of reaction networks. After having applied these methods in the construction, refinement, and verification of our model, we have acquired an extensive mathematical model of pathways relevant for a certain disease.

Using this model (or potentially different model alternatives having a different structure or altered parameter values) one can enter the cycle of Systems Biology as indicated in Figure 5.1. In this cycle a model is used to make a prediction, which is afterwards tested experimentally. Given that the experiments do not support the model predictions, the model has to be refined in order to be in accordance with them. Finally, the next prediction can be made and the cycle starts anew. When this cycle is applied to the identification of drug targets, the generation of hypotheses can be replaced by the directed search for potential treatments. After having tested these treatments experimentally, one has either identified a potent selection of

5.1. DRUG TARGET IDENTIFICATION BY PARAMETER OPTIMISATION

targets or gained additional data to refine the model. In this second part of my thesis I will introduce methods to make predictions on the potential drug treatments, which can afterwards be tested experimentally. The prediction is divided into two parts: the formalisation of the drug target identification problem as a parameter estimation problem and the subsequent application of methods to solve it and the identification of synergistic and antagonistic drug combinations, which requires a slight modification of this formalism. These two topics will be covered in this and the following chapter.

5.1.1 Drug target identification is a parameter estimation problem

Apart from an extensive, validated model of the pathways involved in a disease a description of the so-called “healthy” state of the system is needed in the drug target identification process. The model’s dynamics are supposed to represent the “diseased” state of the system. Thus, the healthy state is what the model should be driven to by the application of external perturbations (compare e.g. [Vera et al., 2007]). Depending on the application, the diseased and the healthy state might describe different behaviours of the investigated pathways. They might describe the overproduction of a certain substance in a human cell and its normal production level or they might describe the working metabolism of a human parasite and a state in which the metabolism has been effectively disrupted.

For the following considerations it is conceptually advantageous, if the healthy state can be expressed by time course data, i.e. values of certain observables at certain time points. In this case, the difference between the dynamic behaviour of the diseased model and the healthy state can be used to construct an objective function equivalent to those used in parameter estimations. Given that the objective of a target identification problem can be formulated in such a way, one can map it to a parameter estimation problem by applying the following steps.

First, one includes potential inhibitors targeting all reactions in a system by various modes-of-action. Second, one declares the concentrations of all added inhibitors as variables for the optimisation. Third, an objective function has to be constructed from the description of the healthy state. This primary objective can be extended by the inclusion of undesired side effects of a potential treatment and it can be accompanied by a secondary objective, e.g. to optimise a treatment for involving a minimal number of drugs. Eventually, one can apply different optimisation methods which results in lists of inhibitor concentrations and values of the objective function(s) prioritising

5.1. DRUG TARGET IDENTIFICATION BY PARAMETER OPTIMISATION

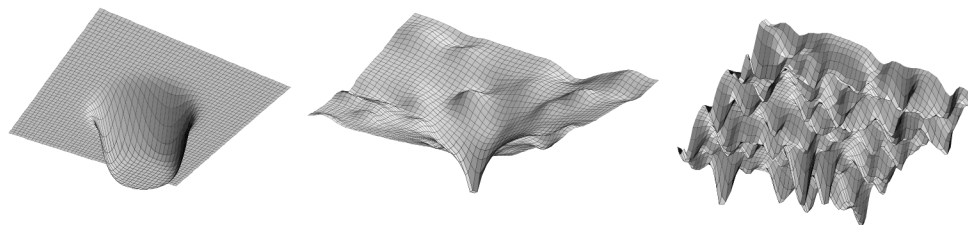


Figure 5.2: Arbitrary landscapes showing different degrees of complexity in the objective function. While the landscape on the left shows a fairly simple optimisation problem that could arise from a simple model and an objective function including only a single time point, the middle and the right landscape could arise from problems including more and more data points in the objective function leading to the fact that they have to be solved using global optimisers to avoid getting stuck in local minima. The figure has been taken from [Feala et al., 2010] and adapted.

the possible treatments.

5.1.1.1 Implications concerning solutions

The shape of the objective function under varying concentrations of the individual drugs can be more or less complex (e.g. as shown in Figure 5.2). It depends on the structural and dynamic complexity of the underlying model, the kinetic details of the drugs' effects on the velocity of the affected reactions, and the number of data points used in the construction of the objective function. The fitness landscape determines how many local optima are available and how these optima look like, e.g. high dimensional subspaces. Thus, in order to solve the drug target identification problem different optimisation strategies have to be followed depending on the model and the objective.

For pharmaceutical applications not all solutions are equally useful. Treatments can in principle involve an arbitrarily large number of inhibitors, but this number should be chosen carefully to find a trade-off between the need for a larger number of interventions to fulfil an objective and the effort in developing the corresponding drugs afterwards. Treating a larger number of targets in parallel has been proposed to be efficient and highly selective [Ágoston et al., 2005], to reliably destroy a system's robustness to perturbations [Lehár et al., 2008], and to slow down the development of drug resistance [Bonhoeffer et al., 1997]. Nevertheless, solutions targeting all reactions in a network (e.g. proposed in Tveito & Lines [Tveito and Lines, 2009]) are not practicable as the development of a treatment involving a large number of

5.1. DRUG TARGET IDENTIFICATION BY PARAMETER OPTIMISATION

drugs requires more effort in drug design stages.

Depending on their complexity different investigated problems can require varying numbers of drugs. Amongst the more simple problem there might be cases in which many local minima with a similar objective value exist. In such cases the inclusion of secondary objectives should be considered to find means to prioritise the solutions. An example of such secondary objective is the minimisation of the amount of drug used in total which is supposed to reduce the probability of encountering unforeseen side effects of the treatment and increase the possibility of achieving a required bioavailability of the final drug.

5.1.2 Parameter identifiability

5.1.2.1 Why does identifiability matter?

Following the above mentioned approach one is able to identify single treatments, which are theoretically able to cure the modelled disease. When setting up an experiment to test a hypothetical treatment in an experiment the problem might arise that certain targets of this treatment are not drug-gable. In the scope of my thesis I tackle this challenge by identifying targets which can replace each other in certain treatments. For this purpose I have applied the concept of parameter identifiability to the drug target identification process.

Identifiability can be applied to any kind of parameter estimation problem and answers the question of how precisely certain parameters can be estimated. More in detail it answers the question, given an optimal solution, can one randomly change the value of one parameter and compensate for it by changing other parameters? In fact, in kinetic models on average only $\frac{1}{4}$ th of the parameters is inferable from experimental data as some model variables cannot be measured with sufficient precision or even not at all [Gutenkunst et al., 2007, Erguler and Stumpf, 2011]. Furthermore, examples exist in which no single parameter can be estimated reliably [Ashyraliyev et al., 2008].

Applied to the drug target identification problem identifiability of a parameter in a solution signifies, whether one target can be replaced by a different target in a certain treatment. Once these alternatives are identified they can be valuable knowledge in the actual selection of targets based on additional biological knowledge.

5.1. DRUG TARGET IDENTIFICATION BY PARAMETER OPTIMISATION

5.1.2.2 Available approaches to find non-identifiable parameters

Non-identifiability of parameters can have two different reasons [Raue et al., 2009]. On the one hand, there is structural non-identifiability, which is a result of the overparametrisation of a model. If the parameter space has a higher dimensionality than the “space of the time courses” of the observables, various parameter combinations can lead to the same time course, rendering some model parameters non-identifiable. These non-identifiabilities can be detected *a priori* from the model using different analytic approaches (e.g. [Pohjanpalo, 1978, Jacquez and Greif, 1985, Ljung and Glad, 1994]) but these approaches do either only deal with linear models or become infeasible for larger model sizes [Bellu et al., 2007, Raue et al., 2011].

On the other hand, there exists practical non-identifiability, which is a result of too sparse data or data being insufficiently accurate. In parameter estimations experimental data restricts the space of “correct” dynamic simulation results, which in turn restricts the space of “correct” parameters. If this restriction is not rigid enough, parameters can remain non-identifiable. This problem can be fixed by including more experimental data, possibly under different conditions, which will render a larger number of parameters inferable [Apgar et al., 2010].

In general, the methods for investigating practical non-identifiabilities work by investigating the shape of the objective function in parameter space around a certain set of parameter values. Practical non-identifiability is therefore a local property of a point in parameter space. As in most cases no closed formula for the objective function exists and since Monte Carlo approaches to approximate it can easily become unfeasible, more simple methods have to be applied to capture its local properties. The shape can for example be approximated by the objective function’s Hessian or using the Fisher information matrix [Vajda et al., 1989, Erguler and Stumpf, 2011]. Such approximations can give hints on local non-identifiabilities and parameter relations, but they are unable to draw a broader picture. To tackle this limitations Hengl *et al.* [Hengl et al., 2007] apply alternating conditional expectation [Breiman and Friedman, 1985] to identify the complex relations which might exist between non-identifiable parameters. Furthermore, Raue *et al.* [Raue et al., 2009] exploit the profile likelihood [Venzon and Moolgavkar, 1988] to determine numerical ranges in which the parameters are unidentifiable. Using these approaches, practical non-identifiabilities can be investigated in more detail.

5.2. FORMALISING THE DRUG TARGET IDENTIFICATION PROBLEM

5.1.3 Host pathogen systems

In most applications the effect of the treatment with a potential drug cannot only be judged by its effect on a single type of cell. When treating a parasitic infection, not only the effect on the parasite but also the effect on the human host has to be considered [Bakker et al., 2000a,b]. The same observation holds for treatments of cancer, in which a drug should selectively harm cells dividing in an uncontrolled manner but no other types of cells [Garrett and Workman, 1999].

Selectivity against a certain cell type can be achieved in two different ways. First, a drug can be designed to specifically bind its target only in the desired cell [Fidock et al., 2004]. Second, its target can be chosen such that the targeted cell type is far less robust to perturbations against it. The first idea requires structural differences in the proteins in between “pathogen” and “host”, which becomes more probable if their amino acid sequences show less homology [Hasan et al., 2006]. The second idea requires differences in the structures or the dynamics of the affected pathways, which can be computed using Systems Biology models of both cell types.

A simplified quantitative picture on how different network dynamics can affect the quality of a target can be given by the application of Metabolic Control Analysis to the field of drug target identification (e.g. [Cascaete et al., 2002, Hornberg et al., 2007, Murabito et al., 2011]). In this context the concept of network-based selectivity has been developed [Bakker et al., 2002]. This quotient compares the derivatives of the steady state values of an observable with respect to an inhibitor’s concentration in between two organisms. Using the selectivity one is able to prioritise drug targets which have a bigger effect in the parasite than in the host. The application of this approach requires dynamic models of the affected pathways in both organisms. However, it compares steady state behaviour under an infinitesimal small perturbation. Applying this concept in my framework it can be extended to non-steady state behaviour and treatment-like perturbations.

5.2 Formalising the drug target identification problem

In the previous section the idea that the drug target identification problem can be mapped to a parameter estimation has been introduced. This idea has been discussed in the literature in various ways. First, the effect of potential drugs on different systems has been studied using different kinds of objective functions comparing the effects [Jackson, 1993, Hornberg et al.,

5.2. FORMALISING THE DRUG TARGET IDENTIFICATION PROBLEM

2005a, Fitzgerald et al., 2006]. Second, drugs acting with different mechanistic modes-of-action on various targets have been included into a pathway and the resulting changes in the steady state behaviour have been compared [Gerber et al., 2008]. Finally, different numerical optimisation methods have been applied to the drug target identification problem [Yang et al., 2008, Tveito and Lines, 2009]. Nevertheless, none of the above mentioned approaches provided a comprehensive and automatable framework and an open source software implementing it.

During my thesis I have collected a set of approaches from the literature and combined them in a framework providing computational support in the prediction of treatments. In the following I will introduce details on how a drug target identification problem can be converted into a parameter estimation problem and how it can be solved. Therefore, I will show how a model should be manipulated and how the variables for the optimisation steps should be selected. Furthermore, I will discuss what results will be gained from the application of different optimisation methods to the problem.

5.2.1 Setting up a parameter estimation to solve the drug target identification problem

5.2.1.1 Identifying and inserting kinetics

The variables of the optimisation problem, which is to be constructed in this section, represent the action of potential drugs on targets in the considered model. In most cases, such variables are not included in a model and have to be added prior to the optimisation. Different approaches how to implement the action of a drug on a certain target exist in the literature, and they differ in the detailedness to which the action is modelled:

1. multiplication of the reaction kinetic with a Boolean variable, which models a knock-out of the corresponding gene (e.g. [Raman et al., 2005, Lee et al., 2009]);
2. multiplication of the kinetics with a non-negative real number, which describes a linear change in the corresponding enzyme's concentration or activity (e.g. [Vera et al., 2007, Yang et al., 2008]);
3. incorporation of inhibition kinetics (e.g. [Gerber et al., 2008, Yang et al., 2008]), which model a drug's mode of action in detail.

5.2. FORMALISING THE DRUG TARGET IDENTIFICATION PROBLEM

Which kind of variables should be used? Most of the drugs approved in recent years are competitive inhibitors [Swinney, 2006]. Competitive inhibition of a certain target results in a dynamic behaviour which is fundamentally different to a gene knock-out or a change in enzyme concentration, as it can be overcome by an accumulation of the substrates, rendering the inhibition useless [Westley and Westley, 1996]. Thus, the effects of a treatment with a potential drug *in vivo* should not be modelled using mechanisms 1 or 2 but rather by detailed inhibition kinetics in order to avoid false positive targets in the prediction.

How are inhibition kinetics inserted into a model? In order to include these detailed inhibition mechanisms, the framework I have developed contains a step in which the model's network structure and kinetics are manipulated. For every selected reaction in the model the kinetics are changed by including possible inhibitors or activators with different modes-of-action (e.g. competitive or non-competitive inhibition). The identification of kinetics in the model and the replacement by inhibition kinetics is done based on numerical comparisons to kinetic formulas from an internal library. This library is based on the Systems Biology Ontology (SBO) [Le Novère, 2006] and has been extended manually to include other popular kinetics, which I have encountered during the analysis of various models. Furthermore, the library can be extended by the user. Given the kinetic formulas in an SBML model this tool adds various possible inhibition kinetics semi-automatically to the library, as shown in subsection B.2.1 in the Appendix.

Apart from regular inhibition/activation kinetics describing the effect of a modifier with a particular mode-of-action, I have introduced superimposed kinetic formulas, which describe the effects of various modifiers at the same time. Single modifications can be added to a kinetic formula by multiplication of certain parameters by a factor representing the binding of the particular modifier (see B.2.1.3 in the Appendix). All these factors for all possible inhibitors/activators can in principle be included into a superimposed inhibition formula. As these introduced factors are equal to 1 when no inhibitor is present, the superimposed kinetic resembles the various inhibition/activation kinetics as long as only one modifier for this reaction is present. Therefore, variables describing the potential action of drugs acting with various modes-of-action on the targets in a given reaction network can be introduced by

- replacing reaction kinetics by their corresponding superimposed inhibition/activation kinetics,
- introducing new variables describing the absolute concentrations of inhibitors acting on different targets with different modes-of-action, and

5.2. FORMALISING THE DRUG TARGET IDENTIFICATION PROBLEM

- adding parameters describing the binding constants of the inhibitors to their targets.

5.2.1.2 Affecting multiple targets with a single drug

The inclusion of new variables as explained above assumes that every reaction can be targeted individually. This might not be the case for every model as e.g. two reactions in the model could be catalysed by the same enzyme. Thus, the inhibition of such reactions with a particular mode-of-action should be described by the same variable.

In order to detect such cases automatically, I propose the following method. First of all, two reactions should be inhibited by the same potential drug if they are catalysed by the same enzyme, i.e. their kinetics are dependent on the same modifier concentration and this modifier's annotations suggest that it is a protein. Judging from an element's annotations I suppose an element to be an enzyme if it is either annotated as a protein, e.g. containing a UniProt annotation [Bairoch et al., 2009], or with an enzyme classification (EC) number. Second, reactions should be inhibited by the same drugs if their annotations or the annotations of their enzymes have a certain similarity. This similarity is determined using the vector-based measure defined in chapter 3. When the (reaction or modifier) elements show a similarity which is higher than a user defined threshold, individual inhibition variables (for all modes-of-action) are replaced by variables representing the parallel inhibitions of both reactions.

5.2.1.3 Preparing the objective function

Now that we have added variables describing the effect of potential drugs to the model, we construct the objective function which should be minimised during the target identification process. This objective function has to be chosen with as much carefulness as the model itself because it will have a large impact on which drug treatments we will regard as optimal.

For the applications described in this work, the model is supposed to describe the diseased state of a system and it is supposed to do that with sufficient accurateness and detail. Opposed to that, the objective function will describe the healthy state, a state to which the model is supposed to be driven by the treatment. Therefore, it has to incorporate information on the molecular consequences of a disease (e.g. an elevated concentration of a certain substance) in order to cure it. Furthermore, general information of the healthy state (e.g. normal concentrations of a few important metabolites) should be incorporated to avoid side effects of the treatment.

5.2. FORMALISING THE DRUG TARGET IDENTIFICATION PROBLEM

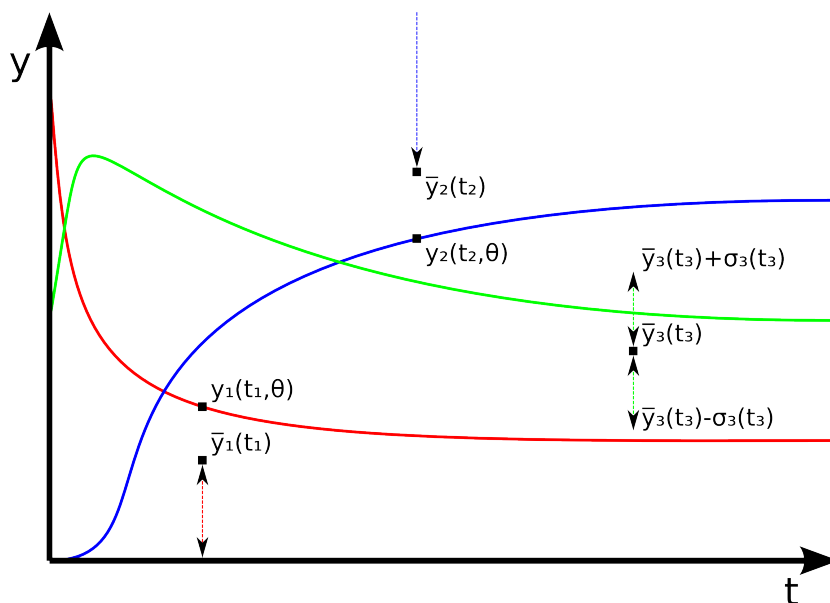


Figure 5.3: Different objectives at different concentration time points. A trajectory of a simulation in the healthy state can be required to fall below a certain threshold, surpass it, or stay within certain boundaries.

Description of the healthy state As a compromise between approaches existing in the literature, I propose to describe the healthy state by a number of concentration time points $\bar{y}_i(t_j)$ for variables in the model. Each time point is to be accompanied by information on how a dynamic simulation of the model should behave under an effective treatment with respect to this point. Given that a substance i has an elevated concentration in the diseased state, this concentration should drop below the point in the healthy state: $y_i(t_j, \theta) < \bar{y}_i(t_j)$, where $y_i(t_j, \theta)$ is a time point of the model simulation run with parameter set θ (e.g. y_i in Figure 5.3). Given that in the diseased state the concentration is too high or not within a certain range, a healthy state should satisfy $y_i(t_j, \theta) > \bar{y}_i(t_j)$ or $\bar{y}_i(t_j) - \sigma_i(t_j) < y_i(t_j, \theta) < \bar{y}_i(t_j) + \sigma_i(t_j)$, respectively (see y_2 and y_3 in Figure 5.3). Using concentration time points to discriminate between diseased and healthy state is supposed to be the most appropriate measure of a treatment's success [Kell, 2006].

With this information, an objective function measuring the difference between a treatment simulation and the healthy state is constructed. In accordance with existing parameter estimation methods, we compute the function

$$\mathcal{X}^2 = \sum_{i,j} \left(\frac{\bar{y}_i(t_j) - y_i(t_j, \theta)}{\sigma_i(t_j)} \right)^2.$$

5.2. FORMALISING THE DRUG TARGET IDENTIFICATION PROBLEM

Assuming that the normalised residuals $(\bar{y}_i(t_j) - y_i(t_j, \theta))/\sigma_i(t_j)$ are independent and standard normally distributed, χ^2 follows a χ^2 distribution. Depending on the objective a “healthy simulation” has to fulfil with respect to each time point, different summands have to be added to the objective function:

- minimisation of a concentration beyond a certain value

$$\chi^2 = \dots + \left(\frac{0 - y_i(t_j, \theta)}{\bar{y}_i(t_j)} \right)^2 + \dots$$

- maximisation of a concentration

$$\chi^2 = \dots + \left(\frac{0 - \frac{1}{y_i(t_j, \theta)}}{\frac{1}{\bar{y}_i(t_j)}} \right)^2 + \dots$$

- keeping a concentration in a range

$$\chi^2 = \dots + \left(\frac{\bar{y}_i(t_j) - y_i(t_j, \theta)}{\sigma_i(t_j)} \right)^2 + \dots$$

Acceptable solutions Instead of performing a statistical test using the χ^2 distribution with the appropriate number of degrees of freedom, we accept all drug target interventions with

$$\chi^2 < 1. \tag{5.1}$$

In this case, we know that all separate aims (minimisation/maximisation of a concentration beyond a certain value or keeping a concentration in a certain range) are fulfilled. For a proof, see section B.2.2.1 in the Appendix.

5.2.2 Different optimisation methods and what solutions to expect

As mentioned beforehand, the objective function can vary in its complexity across different problems. Given a fairly simple objective, one might be able to successfully identify a global optimum using a local optimiser, e.g. Nelder Mead (Simplex) [Nelder and Mead, 1965] or Broyden–Fletcher–Goldfarb–Shanno [Broyden, 1970, Fletcher, 1970, Goldfarb, 1970, Shanno et al., 1970]. In more complex cases many local optima will exist, which requires the employment of global optimisers as Simulated Annealing [Kirkpatrick et al.,

5.2. FORMALISING THE DRUG TARGET IDENTIFICATION PROBLEM

1983] or Genetic Algorithms [Goldberg, 1989] in order to find one (or possibly more) treatment(s).

If these algorithms are applied using the original objective function, their result can in principle involve an unlimited number of drugs applied in parallel. While this provides initial information on whether the objective can be achieved, it does not result in an easily testable hypothesis. Therefore, in most cases one would additionally like to reduce the number of inhibitors used. This aim can be achieved in different ways.

First, one can include the number of used drugs into the objective function such that it is minimised in parallel. When doing so, one has to find a trade-off between both objectives (treatment and reduction of drug number). This can be implemented by including both objectives into the objective function together with a factor weighting them. As an appropriate value for this factor is unknown in the beginning it can be increased stepwise until an optimisation run finds a suitable solution containing a sufficiently small number of drugs. The penalty method [Courant, 1943] or augmented Lagrangian methods [Powell, 1969] are examples of how such an approach can be implemented.

Second, the maximal number of drugs could be included as a hard constraint. This would require the employment of optimisation methods incorporating constraints, e.g. COBYLA [Powell, 1994], but it might be necessary to start the optimisation again and again for an increasing number of used drugs as the algorithm might not be able to find a feasible solution.

Finally, the objective of using a minimal number of drugs can be included by a brute force optimisation on top of the individual drug concentration optimisations [Schulz et al., 2009]. In each step of the “outside” optimisation only selected reactions are allowed to be targeted. The concentrations of drugs against them are then optimised in the “inner” optimisation. Along all of these optimisations, the most favourable treatments, e.g. the ones using the smallest number of drugs in the lowest concentrations, can be identified afterwards. Although the brute force approach is numerically more demanding, it allows for various types of constraints and, given that the inner optimisation finds the best solutions, is guaranteed to find an optimal solution of the “outside” objective. Different authors have proposed similar approaches but solved the “outside” optimisation by non-exhaustive searches [Yang et al., 2008, Calzolari et al., 2008]. These approaches save computational effort at the price of potentially missing optimal treatments. However, this process is easily parallelisable and the time these computational steps require is negligible compared to the total time it takes to develop an effective and safe drug. Thus, there is no need to rely on computational approximations of optimal treatments.

5.2. FORMALISING THE DRUG TARGET IDENTIFICATION PROBLEM

5.2.3 Parameter identifiability

Non-identifiable parameters Given the result of a successful parameter estimation, identifiability analysis can answer the question of how precisely the model parameters are determined by the experimental data. In the application of drug target identification this analysis can answer, whether a change in the concentration of a drug in a certain treatment can be compensated by the change of the concentration of a different drug. Using this analysis we can therefore determine treatment alternatives for cases in which a selected target is not or hardly druggable.

Relation between non-identifiabilities and the objective function

As for general parameter estimations, there exist two different types of non-identifiabilities between potential drugs [Raue et al., 2009]. The first type is structural non-identifiabilities. These are caused by potential drugs having the same effect on all observables in a model. Reasons for such non-identifiabilities are easily detectable cases, e.g.

- when the same reaction is inhibited by inhibitors with a slightly different mode of action,
- when reactions catalysed by isoenzymes are treated in parallel,
- or when a reversible reaction has been broken up into two irreversible reactions and those are treated with one inhibitor and one activator.

As such cases can be avoided by proper selection of treatments tested in parallel, I will not further discuss them in the scope of this thesis.

The second type of non-identifiabilities are practical ones. These arise in between two hypothetical drugs, when both of them result in the same effect on the variables covered by the objective function. This means either of the two drugs can be used to cure the disease and avoid all the side-effects which are currently observed. Therefore, the non-identifiable drugs can be used to serve as treatment alternatives, using either one drug, the other drug, or a combination of both in the final treatment. Alternatively, the simulations under different treatments can be used to find observables being affected in different ways. Biological knowledge on these observables might lead to the identification of further relevant side-effects that should be integrated into the objective function. The successive integration of new knowledge will remove practical non-identifiabilities and will lead to a safer treatment eventually.

5.2. FORMALISING THE DRUG TARGET IDENTIFICATION PROBLEM

Discovering non-identifiabilities Now that I have illuminated why we investigate non-identifiabilities, I will introduce a method to determine them computationally. For this purpose, I formulate identifiability specifically for the drug target identification process.

In the following I will denote a working treatment involving the potential drugs x and y by $T(\{x, y\})$. Knowing the concentrations i_x and i_y of those drugs I will denote a quantitative treatment by $T([i_x, i_y])$, i.e. there exists a drug concentration vector $\theta_{i_x, i_y} = (\theta_1 = 0, \theta_2 = 0, \dots, \theta_x = i_x, \dots, \theta_y = i_y, \dots)$ such that $\mathcal{X}^2(\theta_{i_x, i_y}) < 1$. Let $T([i_{x_1}, \dots, i_{x_n}, i_y])$ and $T([i_{x_1}, \dots, i_{x_n}, i_z])$ be working treatments, then I call y replaceable by z in the condition $\theta_{x_1} = i_{x_1}, \dots, \theta_{x_n} = i_{x_n}$, in short $I(i_y \rightarrow i_z | i_{x_1}, \dots, i_{x_n})$, if

$$\forall_{0 \leq i'_y \leq i_y} \exists i'_z : T([i_{x_1}, \dots, i_{x_n}, i'_y, i'_z]).$$

Furthermore, I call y and z interchangeable in the condition $\theta_{x_1} = i_{x_1}, \dots, \theta_{x_n} = i_{x_n}$, in short $I(i_y \leftrightarrow i_z | i_{x_1}, \dots, i_{x_n})$, if $I(i_y \rightarrow i_z | i_{x_1}, \dots, i_{x_n})$ and $I(i_z \rightarrow i_y | i_{x_1}, \dots, i_{x_n})$. This definition is of biological relevance, as interchangeable drugs in a treatment can replace each other if one of them can only be applied in a lower than required dose or even not at all.

In order to identify all drug interchangeabilities I propose the following algorithm:

```

for  $\mathcal{I}_1 \in \mathcal{P}(\mathbb{I})$  do
  if  $T(\mathcal{I}_1)$  then
    for  $\mathcal{I}_2 \in \mathcal{P}(\mathcal{I}_1)$  do
      for  $\mathcal{I}_3 \in \mathcal{P}(\mathbb{I} \setminus \mathcal{I}_1)$  do
        if  $I(i_{\mathcal{I}_2} \leftrightarrow i_{\mathcal{I}_3} | i_{\mathcal{I}_1})$  then
          report interchangeability
        end if
      end for
    end for
  end if
end for

```

The test for interchangeability of drug combinations is implemented by successively lowering i_y on a logarithmic scale, e.g. $i'_y = i_y \cdot f^{-1}$, $i_y \cdot f^{-2}$, \dots , $i_y \cdot f^{-g}$, and then applying a global optimiser to find a value i'_z for which $T([i'_y, i'_z, i_{\mathcal{I}}])$. In case \mathcal{I}_2 involves more than one drug, i'_y values are varied on a comparable high-dimensional grid. If the optimisation leads to acceptable \mathcal{X}^2 values for all values of i'_y , we assume that $I(i_y \rightarrow i_z | i_{\mathcal{I}})$.

This algorithm explores the complete combinatorial space of drug combinations and has therefore a running time of $\mathcal{O}((2^n)^3)$ with n being the number of possible targets. As this running time is already unacceptable

5.2. FORMALISING THE DRUG TARGET IDENTIFICATION PROBLEM

for identifying interchangeabilities in models of moderate size, this general algorithm has to be speeded up. I propose to decrease the running time by restricting the cardinality of the sets considered in the power sets of inhibitors $\mathcal{P}(\mathbb{I})$ and pruning the combinatorial space of drug combinations by removing trivial results from further considerations. A full list of rules to infer trivial results and an extended version of the algorithm are shown in section B.2.3.3 in the Appendix.

As an example of how the combinatorial space can be pruned, I introduce the principle that the definition of interchangeability requires “minimality” of the working treatments. I define a minimal working treatment by being a set of drugs \mathcal{I} such that $T(\mathcal{I})$ from which no drug could be removed: $\forall_{x \in \mathcal{I}} \overline{T}(\mathcal{I} \setminus \{x\})$, where $\overline{T}(\{x, y\}) \Leftrightarrow \forall_{i_x > 0, i_y > 0} \mathcal{X}^2(\theta_{i_x, i_y} \geq 1)$. The treatments are supposed to be minimal because working treatments can be extended arbitrarily, i.e. given a set of drugs \mathcal{I} in a working treatment being effective at concentrations $i_{\mathcal{I}}$ we have $T(\mathcal{I}) \Rightarrow T(\mathcal{I} + \{x\})$. Therefore, it follows from $T(\mathcal{I})$ that $I(i_x \leftrightarrow i_y | i_{\mathcal{I}})$. According to this idea, \mathcal{I}_1 should be checked for being minimal at the start of the first loop. Using this idea can lead to a significant reduction in computational effort depending on the model investigated.

5.2.4 Host-pathogen interactions

5.2.4.1 Network selectivity in metabolic control analysis

In the methods presented so far I have assumed that side-effects of a potential treatment can be observed within a single model. For many applications this is obviously not possible and one has to observe the effect of a treatment on different models, e.g. the same pathway in a parasite and a host, in order to come up with a more conclusive estimation of a drug’s safety. Given different relevant models, we can simulate the effects of a treatment individually using the aforementioned methods and compare the results afterwards.

The comparison of the results of drug target prioritisations has first been shown in Bakker *et al.* [Bakker et al., 2002]. In this approach, MCA has been used to identify inhibitors leading to the largest response in an observable. For comparing responses to related inhibitors targeting homologue enzymes in between both models, Bakker *et al.* have defined the “network selectivity” of an inhibitor:

$$selectivity = \frac{C_{v_j}^{S_a}(parasite) \cdot \varepsilon_{I/k_i}^{v_j}(parasite)}{C_{v_j}^{S_a}(host) \cdot \varepsilon_{I/k_i}^{v_j}(host)},$$

where S_a is the observable, I is a potential inhibitor with a binding affinity

5.2. FORMALISING THE DRUG TARGET IDENTIFICATION PROBLEM

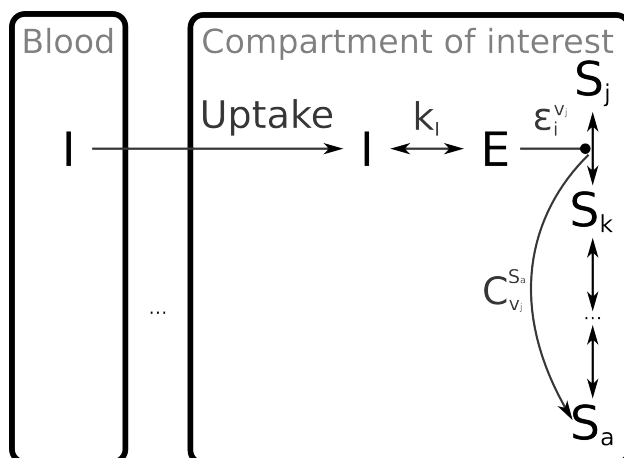


Figure 5.4: Processes contributing to the effect of an administered drug I . The processes include the uptake of the inhibitor to the compartments in which it is supposed to act, the binding of the inhibitor to its target, the local effect of the bound drug, e.g. on the reaction the target is catalysing, and the global effect that is propagated through the network of interest to an observed variable, e.g. the concentration of a certain compound.

of k_i to the enzyme catalysing reaction v_j , C is the control coefficient, and ε is the elasticity (as introduced in chapter 2). Given that the effect of an administered drug can be dissected into four terms (as shown in Figure 5.4)

$$\frac{dS_a/dI(\text{parasite})}{dS_a/dI(\text{host})} = \underbrace{\frac{U(\text{parasite})}{U(\text{host})}}_{\text{uptake}} \cdot \underbrace{\frac{k_i(\text{host})}{k_i(\text{parasite})}}_{\text{structure}} \cdot \underbrace{\frac{\varepsilon_{I/k_i}^{v_j}(\text{parasite})}{\varepsilon_{I/k_i}^{v_j}(\text{host})}}_{\text{elasticity}} \cdot \underbrace{\frac{C_{v_j}^{S_a}(\text{parasite})}{C_{v_j}^{S_a}(\text{host})}}_{\text{control}}$$

the selectivity describes how much effort has to be put into the design of the actual drug. If the potential drug has a high selectivity, i.e. the network effect in the parasite is much higher than in the host, less effort has to be invested into increasing the preferential uptake of the drug into the parasite or the preferential binding of the drug to the parasitic enzyme.

5.2.4.2 Network selectivity in drug target identification

Selectivity for single inhibitors In the scope of my framework, I have to extend the concept of network selectivity, as the effects of potential drugs on an objective are not quantified by a scalar but given by the objective function. Therefore, one is only able to quantify the potency of a hypothetical drug

5.3. RESULTS

by the minimal concentration needed to achieve a certain predefined effect. Given such a desired effect in terms of an objective function (\mathcal{X}^2), I define the required concentration of a drug by

$$\xi_j(\mathcal{X}_p^2) = \min_{I_j} : \mathcal{X}_p^2(I_j \cdot \vec{e}_j) < 1.$$

Depending on the investigated model, this objective function should quantify different biological objectives. When investigating a model of a parasite this can signify the necessary concentration killing it, and when looking at a human model, it can describe the tolerable dose, which we define as the lowest concentration resulting in measurable damage to the host.

Within my extended network selectivity, one again searches for positions in the network which are highly sensitive in the parasite, i.e. they require low doses of a drug, and highly robust in the host. Thus, I define it by

$$selectivity_j = \frac{\xi_j(\mathcal{X}_{host}^2)}{\xi_j(\mathcal{X}_{parasite}^2)}, \quad (5.2)$$

which again assigns favourable drugs are large selectivity. Using this formula targets of single drug treatments can be compared within two different models. An extension of this method to treatments involving multiple drugs and considering more models is given in section B.2.4 in the Appendix.

5.3 Results

5.3.1 Implementation

I have implemented the framework I present in this thesis in the open source Python library libTI2 (<http://sourceforge.net/projects/semanticsbml/>) and a public web tool (semanticsbml.org/TIde). This library combines the aforementioned methods and different optimisers implemented in SciPy [Jones et al., 2001] with fast ODE solvers, which together allows for the rapid computation of potential treatments. The current workflow followed by my library involves the compilation of the considered model including all possible inhibitors in parallel into a binary. This binary is linked to different ODE solvers written in Fortran (i.e. LSODE [Hindmarsh, 1980], LSODA [Petzold, 1983], SEULEX, RADAU [Hairer and Wanner, 1991], EULSIM [Deuffhard, 1985], LIMEX [Deuffhard et al., 1987]), which allows for a fast simulation of the model under different treatments. Being able to compile the model and reuse it in every iteration step provides a significant decrease in computational time compared to other available simulation tools and libraries,

e.g. SciPy, COPASI [Hoops et al., 2006], or the SBML ODE Solver [Machné et al., 2006]. These tools allow for fast simulations, yet they do not offer the possibility to compile a model or they do not provide an interface to reuse the compiled model for further simulations.

5.3.2 Linear pathway

5.3.2.1 Introduction of the model

As a simple example case to prove the usefulness of my approach I analyse the linear pathway which has been investigated in Gerber *et al.* [Gerber et al., 2008]. This pathway consist of a chain of five reactions connecting the metabolites S_1 to S_6 , of which the first and the last are kept at constant concentrations $S_1 = 1, S_6 = 0$. The reactions in the chain have been assigned reversible Michaelis-Menten kinetics (see Eq. B.1) with parameters $V_{\max}^f = K_{mS} = K_{mP} = 1$ and $V_{\max}^r = 0.2$.

In their article Gerber *et al.* investigate changes in the flux through the network with respect to varying positions at which the chain is inhibited, varying modes-of-action, and varying inhibitor concentrations. From this simple model, Gerber *et al.* have predicted some general rules on which drugs would be most successful in a linear pathway. In the following, I reproduce some of their findings within my framework and extend the results to the case of non-infinitesimal inhibitor concentrations. Furthermore, I will show differences in both approaches and I will add new insights.

Details on the construction of the model and the objective function can be found in section C.1 in the Appendix. In contrast to Gerber *et al.* I observe the change in the concentration of the last substrate in the chain, which can here be regarded as equivalent to the change in flux through the network (see section C.1.2).

5.3.2.2 Insights from the application of my framework

As already mentioned, the objective behind the application of potential drugs to this pathway is the reduction of flux through it. Because the pathway is completely linear, the flux is identical to the steady state reaction velocity of any reaction. With the help of MCA, Gerber *et al.* have shown that inhibitors of any reaction in the network reduce the flux through the network with increasing concentration (compare plots in section C.1.3 in the Appendix). Contrarily, increasing activator concentrations increase the flux through the network.

Apart from the concentrations in which a drug should be applied, one

5.3. RESULTS

can ask for the best targets in the chain. For their parametrisation Gerber *et al.* have identified the first reactions to be more effective targets. From the results I have gained, I propose that the best drug target in the network is determined by the equilibrium constants of the reactions in the model. As an easy example, I will investigate different parametrisations in which all reactions have the same parameters and thus the same equilibrium constant ($eq = \frac{V_{\max}^f K_{mP}}{V_{\max}^r K_{mP}}$). If this equilibrium constant is equal to 1, all positions in the chain in principle show comparable results upon inhibition. In the case of non-competitive inhibitors and activators, the results are the same regardless of the inhibitor's target. Opposed to that, uncompetitive inhibitors show better results when targeting the first reactions, while competitive inhibitors prefer the latter (compare Supplementary Figure C.3). Given that the reactions have a larger equilibrium constant (as in [Gerber et al., 2008]), inhibitions of the first reactions show a stronger effect than inhibitions in the latter ones, and vice versa (as shown in Supplementary Figure C.4).

A final relevant question to be answered is the most appropriate mode-of-action for the drug. As proven in section C.1.1, the best type of inhibition for Michaelis-Menten kinetics is determined by the binding constants and the substrate concentrations. Generally, non-competitive inhibitors show the strongest effect, competitive and uncompetitive inhibitors can work comparably well, but they will never work better. To decide whether an inhibitor should rather work in a competitive or an uncompetitive manner, one should investigate whether the following inequality holds in the steady state of the inhibited system:

$$\frac{S}{K_{mS}} + \frac{P}{K_{mP}} < 1. \quad (5.3)$$

If it holds, then the competitive inhibitor achieves the same effect at a lower concentration than the uncompetitive inhibitor. These results are in agreement with Figure 3 from [Gerber et al., 2008] and they are supported by their numerical simulations in section C.1.3.

Integrating these simple rules, one can explain the differences in the response to various inhibitors as the ones shown in Supplementary Figure C.5. In this network, all parameters have been set to 100 and I again try to reduce the flux through the chain. Since the equilibrium constant is 1 for all reactions, all non-competitive inhibitions work equally well and they show the strongest effect. Competitive inhibitors work far better than uncompetitive inhibitors as the left side in Eq. 5.3 is by far smaller than one. For competitive inhibitors latter positions in the chain seem to work better, even though the position of the competitive inhibitor only has a very small effect on the flux inhibition. This can be explained by the fact that substrate concentrations

are decreasing along the chain, which leads to the effect that more inhibitor is bound to the enzyme when later reactions are targeted. Uncompetitive inhibitors show the contrary behaviour. Here, the first positions in the chain work better as the inhibitor binds to the enzyme-substrate-complex, which is more abundant for higher substrate concentrations.

5.3.3 Glycolysis in *Trypanosoma brucei*

5.3.3.1 Prioritisation of drug targets by necessary effective inhibitor concentrations

As a biological application for my framework I have investigated the glycolysis in *Trypanosoma brucei* in two different mathematical descriptions [Albert et al., 2005, Achcar et al., 2012]. The stoichiometry of both models is depicted in Figures C.6 and C.7 in the Appendix. Prior results have shown that the glycolysis is an essential pathway for the bloodstream form of the pathogen and that a reduction in flux by 50% results in a serious growth deficiency [Albert et al., 2005]. Earlier descriptions of the pathway have already been analysed for potent drug targets as discussed in section 1.5.5 but the most recent versions have only been investigated on the basis of the control of individual reactions over the glycolytic flux. In the following I will therefore prioritise drug targets with my method, i.e. inhibiting the glycolytic flux by 50% and rating targets and inhibition mechanisms by the fact which requires the least amount of drug to achieve this goal. Mathematical details on the implementation of the objective function used are described in section C.2.1 in the Appendix.

The results obtained with my approach are in good agreement with prior results (e.g. [Schulz et al., 2009]) (see Table C.1 for detailed values). In case one searches for targets of non-competitive inhibitors the most favourable enzymes are THT, the pyruvate transporter (PT), PGM, GAPDH, GPDH, ENO, and ALD. If one allows only uncompetitive inhibitors, this target set reduces to TH, PGM, GPDH, GAPDH, ENO, and ALD. For competitive inhibitors only GAPDH and THT prove to be valuable targets as the inhibitor concentration for GPDH and the hexokinase (HK) needs to be 2.5 orders of magnitude higher than its K_I value.

5.3.3.2 What determines the quality of a target

An important question to raise is which factors determinate the quality of the targets in the models. While prior investigations attribute the large control of the THT to the difference in the K_m values for glucose of the THT and the

5.3. RESULTS

HK, I claim that my target prioritisation is mainly determined by the V_{max} values in the enzyme kinetics. This claim can be supported by the following calculations. For each reaction I compute the desired flux v_d . This flux should be defined as the maximal flux through the reaction with which the system can be driven into the “healthy” state. In the case of the glycolysis this can simply be defined as half of the flux in the “diseased” state. Then I compute the needed effective concentration of a non-competitive inhibitor to reduce the maximal enzyme capacity to v_d by solving $v_d = \frac{1}{1+\frac{I}{K_I}}V_{max}$ for the inhibitor concentration and compare this value to the necessary concentration from simulations (concentrations leading to $\mathcal{X}^2 = 1$). The degree in which both effective concentrations differ signify how much the effect of the simulated inhibition can be attributed to the reduced enzyme capacity and how much to changes in the rest of the network (e.g. the change in the metabolite concentrations of its substrates and products).

Setting the desired flux v_d for enzymes in the upper part of the glycolysis to $52 \frac{nmol}{min \cdot mg \text{ protein}}$ and in the lower part to 87, the effective inhibitor concentrations from both calculations, the simulations and the prediction via the enzyme capacity reduction, largely correspond to each other. Only PFK and GAPDH show a significantly smaller value for the necessary inhibitor concentration in simulations than predicted from the enzyme capacity. Therefore, the effect of non-competitive inhibitors on both targets cannot only be explained by the V_{max} values of the enzymes.

In order to test the hypothesis that for most enzymes the maximal capacity determines the potency of the corresponding non-competitive inhibitors, I have overexpressed enzymes in the model 10 fold by multiplying the V_{max} values with the factor 10 and run the drug target identification process again. For most enzymes the necessary inhibitor concentrations increased almost 10 fold, except for the PFK where it only increased marginally. Simulations of the inhibition of PFK show that the inhibition leads to an accumulation of upstream substrates and the production of glycosomal ADP (adenosine diphosphate). This accumulation drives the lower glycolysis to a point at which it runs out of substrates and the system is running into a state of depletion of glycolytic intermediates.

5.3.3.3 Considering side effects in human cells

The development of a drug against *T. brucei* should not only be focused on the efficacy of the treatment against the pathogen. In order to assess the safety of a potential treatment as well, the effects on the human host have to be considered. Therefore, I first look for models describing glycolysis in human cells and then use the network selectivity measure defined by Equation

5.3. RESULTS

Table 5.1: Network selectivities between targets in the Achcar and the Holzhütter model for different modes-of-action. Higher selectivities denote that an inhibition is effective in trypanosomes and shows almost no effect in erythrocytes.

| | non-competitive | uncompetitive | competitive for | |
|-------|-----------------|---------------|-----------------|----------|
| | | | substrate | cofactor |
| ALD | 34.7 | 609 | 0.226 | |
| GAPDH | 62.7 | 36.6 | 110 | 12.9 |
| THT | 14.7 | 19.1 | 9.75 | |
| PGM | 6.72 | | | |
| PGI | 3.21 | | | |
| PGK | 1.79 | 11.4 | 0.0155 | 0.489 |
| PFK | 0.315 | 8.25 | 0.0195 | 0.00603 |

5.2 to look for targets which have a less severe effect in the host than in the parasite.

A search for glycolysis models with semanticSBML (see Figure C.8 in the Appendix) revealed the Holzhütter model of carbon metabolism in human erythrocytes [Holzhütter, 2004] (BIOMD0000000070) as a potential candidate to evaluate drug safety. This model shares a comparably large number of compounds and homologue reaction with the trypanosomal glycolysis. Erythrocytes are of particular interest in this context as they are the most abundant cells in the blood, the place of residence of trypanosomes in the first stage. Furthermore, the Holzhütter model has already been used to assess the safety of potential drugs with the network selectivity using an older version of the trypanosomal glycolysis [Bakker et al., 1999].

Using network selectivity, we compare targets by the effective inhibitor concentrations needed to reduce the flux through trypanosomal glycolysis by 50% and the flux through the erythrocyte counterpart by at most 5%. For details on the objective function used, the reader is referred to section C.3.1 in the Appendix.

The network selectivities depicted in Table C.3 in the Appendix show basically no differences when the Holzhütter model is compared to either the Albert [Albert et al., 2005] or the Achcar model [Achcar et al., 2012]. A small set of selectivities is shown in Table 5.1. Values bigger than one arise when the effective inhibitor concentration needed for inhibition of trypanosomal glycolysis leads to almost no disturbance in the erythrocytes. The most promising targets in this analysis are GAPDH and THT as inhibitors with various modes-of-action against these targets have a high selectivity. ALD, PGM, PGI, and PGK have a high selectivity if targeted in a non-competitive

5.3. RESULTS

manner. Unfortunately, PFK, which has been identified as an interesting target beforehand, shows a comparably bad selectivity if it is not targeted by an uncompetitive inhibitor.

Further targets, which are not included in Table 5.1, can be of interest as well. For drugs targeting TPI and PT the calculated selectivity is infinity, because the inhibitors show no effect in erythrocytes. Furthermore, the enzymes TAO and GPDH are not present in red blood cells and might also be of potential interest [Fessas et al., 1980].

The results presented here largely correspond with the drug target ranking by Bakker *et al.* [Bakker et al., 1999] but give a more distinctive picture on the preferable targets with respect to different modes-of-action.

5.3.4 AA pathway

As a second biologically relevant example, I investigate possible drug targets in the arachidonic acid pathway in humans. This pathway is responsible for the production of various eicosanoids, which play a role as mediators of pain, fever, and inflammations. As an objective, the concentrations of these mediators should be reduced as it happens in response to anti inflammatory drugs like aspirin. At the same time known side effects like gastrointestinal bleeding or strokes, which are caused by an imbalance of other eicosanoids, should be reduced.

5.3.4.1 Preparation of model

Refining the model In section 4.3.4 in the previous chapter I have constructed five different models of the arachidonic acid pathway in three different cell types: granulocytes, endothelial cells, and platelets [Yang et al., 2008]. Because of lacking references for the number of cells and the cellular volume in the publication of Yang *et al.*, which are necessary to construct the aforementioned equations, these numbers are recalculated from reported data in Table C.4 in the Appendix. As a final refinement of the models I include new equations to calculate the area under the curve (AUC) of the observed eicosanoids LTB₄, PGE₂, PGI₂, and TXA₂. This AUC is supposed to be a measure of the activity of downstream receptors.

Setting up the objective function The objective function we try to minimise with the help of different drugs is the sum of the following parts:

- reduction of the AUC of LTB₄ to 10%,
- reduction of the AUC of PGE₂ to 10%,

- keeping the ratio of the AUCs for PGI₂ and TXA₂ within 20% of its value in the untreated state.

The full function is given in section C.4.1 in the Appendix.

5.3.4.2 Prior results

Solutions of Yang *et al.* So far the model has already been investigated in Yang *et al.* [Yang *et al.*, 2008] with a similar method and a similar objective function. In their paper, they first classify inhibitors by having a “robust” influence on the reduction of LTB₄ and PGE₂ production, i.e. they perform a global optimisation with enzymatic activities as the variables and afterwards rank the enzymes by having a large median over the standard deviation of enzyme activity changes along all accepted solutions. With this step they identify PLA₂, LTA4H, 5-LOX, PGES, and COX-2 as potentially robust targets for inhibitors.

In a second step, the authors add competitive inhibitors against COX-1 and the aforementioned 5 targets to the model and “cure” the model, i.e. reduce LTB₄ and PGE₂ production whilst keeping the PGI₂ to TXA₂ ratio constant, by optimising their effective concentrations. This step is again performed multiple times and statistics over the results are made, which have led to the following conclusions:

- Treatments, which reduce the production of LTB₄ and PGE₂ whilst reducing side-effects, exist and they can involve from 2 up to 6 inhibitors.
- For treatments involving many inhibitors the objective function is less sensitive to variations in the inhibitors concentrations.
- For almost all treatments involving the two cyclooxygenases, COX-1 and COX-2 have to be inhibited in a fixed ratio. This does not hold if PLA₂ is inhibited as well.

Finally, they suggested two solutions how the system should be treated with competitive inhibitors. Their first suggestion involved inhibitions of LTA4H in combination with PGES or the cyclooxygenases. With this treatment it is in principle possible to successfully cure the system targeting only two different enzymes (cyclooxygenases are counted as a single target because of a high structural similarity). The second solution involves the targets PLA₂ and COX in combination with either LTA4H or 5-LOX. This three target combination has the advantage that it allows larger dosing ranges and therefore provides a potentially safer treatment.

5.3. RESULTS

Criticism In the first step of their heuristic Yang *et al.* detect 5 potential targets of inhibitors, which reduce LTB_4 and PGE_2 production. For all further considerations other targets are disregarded because of their lack of robustness or because they would require activators to achieve the desired effects. To this set of targets the authors manually add COX-1. Using this set of six inhibitors the authors then try to fulfil the objective of reducing the eicosanoid production and keeping the PGI_2 - TXA_2 ratio constant, in order to avoid side-effects of the treatment. However, the inclusion of COX-1 is done without any algorithmic justification. Nevertheless, COX-1 does appear in the majority of their final solutions and it should be regarded as a relevant target. Thus, in order to regard further, potentially relevant targets in the treatment identification, I will run my analyses on the full set of targets in the models.

In their analyses in the second optimisation step Yang *et al.* randomly produce various treatments involving the 6 aforementioned inhibitors. These treatments are afterwards clustered and rated by whether they appear along different parametrisations of the model and how sensitive the objective function is to changes in the inhibitor concentrations. Based on these results the authors suggest a number of treatments. However, the final treatment suggestions are not those solutions performing optimal to their own criteria.

For my analyses I will apply the straightforward criterion of achieving an objective with (a) the smallest numbers of inhibitors possible, such that the development of the treatment does not become too difficult, and (b) the lowest inhibitor concentrations, to avoid unforeseen side-effects of the individual drugs. Furthermore, I will replace the Monte-Carlo optimisation by a brute-force approach in order to solidly and reproducibly identify possible treatments.

5.3.4.3 Results

Single target solutions As a first analysis, I performed linear brute-force scans of the effects of single inhibitors and activators on the objective function. Effective concentrations have been varied in the range from 10^{-1} to 10^6 in 70 logarithmic steps for all possible inhibitors and activators in the model. Furthermore, this step has been repeated for all 5 published parametrisations of the model. The results of this step suggest that no practicable single target solutions are available. An activator of 12-LOX and a non-competitive inhibitor of PLA_2 have been able to work for one of the parametrisations, but in both cases their safe concentration ranges were relatively small.

Instead of asking for a single inhibitor or activator that can achieve all three sub-objectives, LTB_4 and PGE_2 reduction and balancing PGI_2 to TXA_2

levels, I have then investigated how each of these objectives is affected by single inhibitors. The results of this analysis are shown in Tables C.5 and C.6 and Figure C.11 in the Appendix. The lists of potential inhibitors of LTB_4 and PGE_2 production overlap in inhibitors for PLA_2 and activators for phospholipid hydroperoxide glutathione peroxidase (PHGPx), 15-LOX, and 12-LOX, making them the only possible single target solutions. However, dependent on the chosen parametrisation of the model, these inhibitors and activators lead to side-effects, which renders them unable to fulfil all three given objectives. This effect on the objective function is visualised for the potential single target solutions in Figure C.12 in the Appendix.

It should be noted that the potential single targets are not part of the reaction chains leading to the production of LTB_4 and PGE_2 . Instead, the targets are either upstream of both production chains (inhibitors of PLA_2), leading to a reduced activity of the whole system, or in other reaction chains (activators of 12-LOX, 15-LOX, and PHGPx), which divert AA away from LTB_4 and PGE_2 production. Their common action in the pathway is thus to remove the available AA, which can afterwards not be metabolised to LTB_4 and PGE_2 .

Dual target solutions Also for the two target solutions I investigate the effects of various inhibitors and activators by changing the concentrations on a grid in the range 10^{-1} to 10^6 using 70 logarithmic steps in each direction. From these simulations I select those inhibitors pairs, which are potentially able to reduce the \mathcal{X}^2 value to 1 or less. Furthermore, for all accepted inhibitor pairs I compared the size of areas in parameter space, in which the treatment would work. Inhibitor combinations having larger areas in space can be considered to be more useful, as the metabolism of the applied drug needs to be taken into consideration when dosing strategies are developed. A larger area can lead to a treatment where fewer development effort has to be put into the final formulation in order to increase the drug's bioavailability over time.

The accepted inhibitor combinations computed in the step fall into two different classes. First, there are pairs of drugs from which each one blocks either LTB_4 or PGE_2 production. Second, other pairs consist of one drug from the aforementioned single targets and a second drug which is compensating the first ones shortcomings, i.e. incomplete suppression of LTB_4 or PGE_2 production or PGI_2 - TXA_2 ratio changes.

The first group includes inhibitors of PGES which can be used in combination with an activator of CYP4F3 or an inhibitor of LTA4H or 5-LOX. Of the three targets inhibiting LTB_4 production, the noncompetitive inhibitor of

5.3. RESULTS

LTA4H and the activator of CYP4F3 work best. These two targets work for all parametrisations and they do so at the lowest effective inhibitor concentrations. However, along different parametrisations the required concentrations differ largely, as already indicated in Tables C.5 and C.6. This fact renders further experiments and model improvements necessary to prioritise between the two targets. Inhibiting 5-LOX, the third target in the LTB₄ production, only worked for four out of five parametrisations. Nevertheless, the required concentrations were in the same range as for LTA4H and CYP4F3. Along all of these 3 target pairs the objective function decreased monotonically with increasing inhibitor concentrations. Therefore, a higher-than-required drug dose does not lead to side-effects, which makes this target combinations highly preferable for treatment.

The second group of targets consists of pairs, in which one target was already mentioned in the single targets list and the other one compensates for the first drug's shortcomings. This combination makes the treatment less sensitive to the parametrisation of the model, as the treatments work for more than one parametrisation, and it increases the dosing range along which the treatments work. To investigate these pairs in more detail, I will have a closer look on pairs involving the inhibition of PLA₂. The inhibition of PLA₂ leads to a decreased production of LTB₄, PGE₂, and PGI₂, but it has no effect on TXA₂ levels. This stems from the fact that platelet cells in the model contain a large pool of free arachidonic acid that is converted to TXA₂ afterwards. Therefore, targets being selected together with inhibitors of PLA₂ are responsible for increasing the PGI₂-TXA₂ ratio, as shown in Figure C.13 in the Appendix, namely inhibitors of TXAS and the PGI₂ to PGF₂ transformation and activators of PGIS, TXA₂ to TXB₂ transformation, and 12-LOX. However, as shown in Figure C.13, the objective function is fairly sensitive to changes in the concentration of the second inhibitors (not to changes in the inhibitor of PLA₂), thus reducing the applicability of these solutions. Similar observations also hold for other pairs of the second group, which makes the first group generally more preferable.

Higher order solutions and target non-identifiabilities For evaluating treatments involving more than 2 drugs in parallel, brute-force approaches cannot be employed anymore. Using a similar approach as for single and dual targets would require more than 100 billion simulations to be performed. Therefore, I will try to assess higher order solutions using a different approach.

So far, I have already discovered two basic ways in how a treatment can work in order to drive the model to the healthy state. One can either re-

duce the LTB_4 and PGE_2 concentrations in each production subpathway individually or one can reduce the availability of arachidonic acid to these subpathways. When pursuing the latter idea with for example an inhibitor of PLA_2 , the treatment also has to take care of secondary problems as the restoration of the original PGI_2 - TXA_2 ratio. If my understanding of these treatment actions is true in general, then it should for example not matter whether one inhibits 5-LOX or LTA4H in combination with PGES as a realisation of the first idea. In particular, one should also be able to inhibit both enzymes, 5-LOX and LTA4H, in combination with PGES. Furthermore, such a treatment should work for varying concentrations of 5-LOX and LTA4H. In my framework this concept can be stated by the hypothesis that there exists a non-identifiability between 5-LOX and LTA4H when PGES is simultaneously inhibited.

Given the algorithm proposed in the methods section of this chapter, I indeed confirmed these non-identifiabilities, which supports my conceptualisation of how the treatments work. Figure C.14 in the Appendix shows how the non-competitive inhibitors of 5-LOX and LTA4H and the non-essential activator of CYP4F3 can replace each other when applied in combination with a non-competitive inhibitor of PGES. The non-identifiabilities imply in particular that also unions of the non-identifiable target sets will provide suitable solutions. With the help of this idea it is possible to build higher order solutions from lower order solutions and the non-identifiabilities, which allows to explain the various kinds of treatments identified by Yang *et al.* [Yang *et al.*, 2008, Table 2].

The fact that inhibitors of LTB_4 production are non-identifiable in conjunction with a non-competitive inhibitor of PGES, which reduces PGE_2 production, suggests that non-identifiabilities might also exist in the PGE_2 production subpathway. Testing this hypothesis I have found non-identifiabilities in between a combination of drugs targeting COX-1 and COX-2 and an inhibitor of PGES, when these are applied with a drug targeting either CYP4F3, LTA4H, or 5-LOX. However, treatments involving COX inhibitors and inhibitors of LTB_4 production would not have been suggested on the basis of this model and the used objective because of the following reasons: First, the non-identifiability did only exist across few model parametrisations. Second, the solution is very sensitive to changes in the concentration of the COX inhibitors. Third, the shape of the objective function in inhibitor space varied heavily across parametrisations, leaving the degree to which an inhibitor should bind one isoform of COX over the other highly dependent on the model parameters.

As a further example of non-identifiabilities in the second group of treatments, I have confirmed the non-identifiability of the PGI_2 - TXA_2 ratio in-

5.4. DISCUSSION

creasers in the context of the non-competitive inhibition of PLA₂. A visualisation of these non-identifiabilities is presented in Figure C.15 in the Appendix. However, as already observed, these solutions are not of practical importance as the objective function is quite sensitive to changes in the non-identifiable targets, which reduces their value for practical treatments.

5.4 Discussion

5.4.1 What has been achieved in this chapter

In this chapter I have introduced a framework that can be used to identify efficacious and safe drug targets on the basis of reaction networks described in terms of ordinary differential equations. The presented framework comprises two parts: First, it manipulates a given kinetic model by inserting various inhibition kinetics in a way that allows for the inclusion of inhibitors with different modes of action into the same formulas and, therefore, into the same model. Second, it formulates the identification of drug targets as an optimisation problem. This has already been done in different contexts, e.g. [Yang et al., 2008, Tveito and Lines, 2009], but to my knowledge it has not been formalised strongly. This formalisation allows for the integration of previous approaches and enables its universal reusability.

The examples in this chapter have demonstrated the usability of my framework. All analyses have been performed using my publicly available web tool, which has thereby proven its analytical versatility. For the different analyses, I have used diverse biological objectives. The fact that these have easily been included into the objective function used for the optimisations underlines the extensibility of the presented approach. Furthermore, biological results agree in general with prior knowledge, supporting the reliability of my predictions. Nevertheless, the presented results are more extensive than those produced by previous approaches and, through the inclusion of additional objectives, are more focused on the practical value of the predictions to the pharma industry.

5.4.2 Comparison to other approaches

5.4.2.1 Predicting valuable drug targets

The results that I have gained using my framework have already been compared to the most conclusive results presented in the literature. Nevertheless, they can also be compared to results provided by other approaches, which do not require a dynamic description of the pathway of interest.

Stoichiometric approaches Using purely stoichiometric approaches on the basis of KEGG pathway maps [Kanehisa et al., 2008] the set of potential results is comparably large. With the help of choke point analysis [Yeh et al., 2004] on the complete metabolic map, one can identify reactions that are the way to produce or consume a certain metabolite. This analysis returns all enzymes in the extended glycolysis upstream of PEP except for GPI and TPI as potential targets.¹

If one disregards reactions catalysed by isoenzymes, because a potential inhibitor might need to target different proteins, one loses THT, HK, GAPDH, PGK and ENO. Interestingly, the analysis of the kinetic model has shown that these enzymes are comparably good targets because of their limiting capacity. This coincidence could be explained by two different ideas [Jackson, 2007]. On the one hand, gene duplication, which often occurs in tandem arrays (HK, GAPDH, PGK), might have led to an increase in the enzymes' concentrations resulting in a higher flux through glycolysis and a selectional advantage. On the other hand, trypanosomes do express some isoenzymes in different environments or transport them to different compartments, which might act as a kind of control mechanism for the metabolism.

Flux balance analysis related approaches The application of flux balance analysis based approaches to the Albert model leads to similar, yet different results. Given the constraints that glucose is consumed by the system and the same amount of pyruvate has to be produced by glycolysis², two different fluxes are possible: The first flux activates the complete glycolysis and the glycerol production. Here, the lower part of the glycolysis operates at half its maximal speed as the NAD/NADH balance has to be restored through glycerol production, which consumes half of the triose phosphates. In the second flux, the complete glycolysis operates at full speed, while for each mole of pyruvate produced by the system the cycle of GPDH, the antiporter, and the glycerol-3-phosphate oxidase (GPO) has to restore one molecule of NAD. These two behaviours are the two elementary modes which consume glucose and produce pyruvate [Schuster et al., 2000]. In case a drug disrupts both of these extreme fluxes, the parasite cannot convert glucose to pyruvate and therefore cannot grow. Thus, enzymes being active in both modes, comprising all glycolytic enzymes including GPDH and excluding TPI, should be considered essential enzymes from this analysis and thus as potential targets.

¹http://www.genome.jp/kegg-bin/show_pathway?org_name=tbr&mapno=01100

²This idea is consistent with maintaining more than half of the original glycolytic flux under aerobic conditions, which has also been used as the objective function for the analysis in my framework

5.4. DISCUSSION

Table 5.2: Lowest sequence identities amongst human and trypanosomal isoenzymes. The results have been obtained by running BLASTp on trypanosomal sequences from KEGG comparing them to the RefSeq sequence database.

| Enzyme | Percent sequence identity |
|--------|---------------------------|
| TAO | n.a. |
| PGM | 25 |
| GPDH | 28 |
| HK | 37 |
| PFK | 46 |
| GK | 46 |
| ALD | 49 |
| PK | 51 |
| TPI | 54 |
| PGK | 56 |
| GPI | 57 |
| ENO | 63 |
| GAPDH | 65 |

Sequence based approaches Many available approaches integrate protein sequence or structure information into the target prediction. In order to compare results obtained with such an approach to results obtained by others I have computed the protein sequence similarity between homologue enzymes from the pathogen or the host in Table 5.2. This analysis has been performed by blasting [Altschul et al., 1990]³ trypanosomal protein sequences from KEGG against the RefSeq sequence database [Pruitt et al., 2000] and choosing the highest sequence similarity between isoenzymes.

Comparing results of selected available methods In general, approaches for target identification relying solely on structural information can only result in binary decisions whether an enzyme might be a good target or not. For the trypanosomal glycolysis these approaches do not lead to a restriction in target space as almost all glycolytic enzymes are necessary to maintain a flux through the pathway and are therefore chokepoints. Amongst the results of stoichiometric approaches only marginal differences exists. Compared to chokepoint analysis, FBA based methods regard the network as a whole and do not only use the local connectivity. Thus, their results differ in small details: FBA does not predict the GPO to be essential.

³<http://blast.ncbi.nlm.nih.gov/Blast.cgi>

This is in accordance with my predictions and the fact that trypanosomes can (at least for a short time) survive anaerobic conditions.

Approaches using kinetic models are able to quantify how good targets are. This information is useful in the construction of actual drugs, as more potent targets do not necessarily require drugs to be highly selective or to have a high bioavailability. In cases in which a mathematical model for relevant pathways is at hand, approaches based on dynamic simulations should be preferred over metabolic control analysis, as they simulate finite changes in inhibitor concentrations, which can be fundamentally different to infinitesimally small perturbations [Schulz et al., 2009].

Information on the sequence and structure similarity of enzymes from host and parasite can be seen as information, which is independent from results gained by network selectivity. The bigger the difference in the enzymes' structures, the larger are the odds of a drug being selective against one of them. Therefore, it can be assumed that selective inhibitors against trypanosomal enzymes in glycolysis can in principle be found. Opposed to that, the network selectivity describes how selective a drug needs to be in order to work. The ordering of the targets in Table 5.2 reflects this independence: TAO, whose inhibition does only show an effect in combination with glycerol, ranks highest while GAPDH, which is supposed to be an efficacious and safe target, ranks last.

In general, structural information can be used to complement selectivity information, as targets with an unfavourable selectivity could still be of interest if the structures of the homologues of the targeted enzymes are fundamentally different. Nevertheless, this would complicate the design of the actual drug.

5.4.2.2 Identifiability

In contrast to the MTOI method [Yang et al., 2008], my identification of treatments is largely based on brute-force optimisation. This leads to a more thorough investigation of treatments but it also makes the investigation of higher order combinations unfeasible. Identifiability analysis, however, provides means to judge which higher order combinations would work and leads to a better understanding of the mechanisms behind successful treatments than MTOI. This is reflected by the fact that the basis solutions, the ones that can explain most of the possible solutions, in the paper of Yang *et al.* only covered 61% of their stochastically determined treatments. Basis solutions identified through non-identifiabilities, in contrast, covered 87% of them (see Figure C.7 in the Appendix). Thus, these solutions provide a more complete picture of the factors needed for a successful treatment, which might be vital

5.4. DISCUSSION

to gain an understanding of how a system can be manipulated effectively.

Methodologically, my approach for the identifiability analysis can be related to the determination of parameter non-identifiabilities via the profile likelihood [Raue et al., 2009]. This approach works by varying one parameter within a range that should be investigated and then optimising other parameters in order to minimise the objective function. The resulting values of the objective function are then observed in relation to the changes in the first parameter’s value. Following this approach one can determine ranges in which a parameter could be varied without necessarily increasing the value of the objective function.

In principle, I perform the same analysis for drug combinations instead of general parameters. However, I start the profile likelihood investigation from the “borders” of the parameter space, as initially most of the drug concentrations are equal to zero. Furthermore, I extend the concept to identify certain non-identifiabilities by only allowing selected drug concentrations to compensate for a decrease in one or more drugs within a treatment.

5.4.3 Biologically relevant results

5.4.3.1 Glycolysis in *Trypanosoma brucei*

General insights As a general conclusion obtained from the results in this chapter I would like to underline that different methodological approaches or objectives might lead to different prioritisations of drug targets. A prioritisation obtained through protein sequence analysis is completely unrelated to a prioritisation using MCA-based methods. In particular, this conclusion underlines that targets which are selected for the efficacy of a drug against them are not necessarily safe as well. Thus, the explicit evaluation of a drug’s safety *in silico*, e.g. by means of network selectivity, should be regarded as necessary.

Potential targets Based on the analysis of the trypanosomal glycolysis in this chapter I propose that further efforts in the development of trypanocidal drugs should focus on the targets GAPDH and THT. Both targets can in principle be inhibited by inhibitors with different modes-of-action (competitive, uncompetitive, and non-competitive) and simulations suggest that these inhibitors will be efficacious and safe. The trypanosomal GAPDH shows the highest sequence similarity when compared to human counterparts and might therefore have been rejected by other approaches. However, as its network selectivity is quite high, efficacious inhibitors of it do not need to be selective for the parasitic enzyme in order to be safe.

Apart from these results, further research could be put into the modelling of the PFK reaction, as simulations have shown interesting effects upon its inhibition. The idea of PFK as an interesting target has been supported by RNA interference experiments, in which the system's response to a reduction in the PFK level is much stronger than predicted from the Albert model [Albert et al., 2005]. Therefore, in order to predict the consequences of PFK inhibition more reliably the models should be refined. One potential idea is the inclusion of a control of the PFK activity by AMP [Cronin and Tipton, 1985], which might reduce the models' abrupt response to PFK inhibition. A second idea is the inclusion of protein translation into the model. Albert *et al.* have identified that a change in PFK activity leads to a change in other enzymes' activities [Albert et al., 2005]. This control of enzymatic activity cannot happen at the level of transcriptional control as this is absent in trypanosomes. In the parasite almost all genes are expressed to a comparable extent and the protein levels seem to be controlled by the 3'UTRs (untranslated regions) of the mRNAs [Clayton, 2002]. Via this mechanism it is for example determined which TXT, ALD, and PGK is expressed in the pathogen depending on the current environment [Hotz et al., 1995].

Another interesting target is the GPDH. Simulations suggest that a drug against it will be highly efficient, but its safety could not be assessed by comparisons to the erythrocyte metabolism, as red blood cells do not contain this enzyme [Fessas et al., 1980]. Thus, further *in silico* validation of this target requires a mathematical model of a relevant human cell type expressing GPDH.

Potential resistances All results which have been gained in this chapter depend on the fact that glycolysis is an essential pathway for bloodstream *T. brucei*. However, experimental results have shown that the glycolysis can become dispensable when the parasite is gradually transferred to a medium with a different energy source [Drew et al., 2003]. Thus, the potential development of resistances against a drug targeting glycolysis will need to be considered and possibly monitored after the treatment is in use.

5.4.3.2 Arachidonic acid pathway

General insights The arachidonic acid pathway, as it is modelled by Yang *et al.*, is a complex system. The various subpathways branching from arachidonic acid and their metabolites interfere with each other through feedbacks leaving the system's detailed dynamic response highly unpredictable. This complexity makes the system suitable for modelling with ODEs, which in

5.4. DISCUSSION

turn opens up the possibility of investigating possible drug targets using my method.

Potential targets Treatment identification, as performed in the results section of this chapter, has suggested different solutions. The most promising solution is the inhibition of PGES in conjunction with inhibitors of LTB₄ production, i.e. non-competitive inhibitors of LTA4H or 5-LOX or an activator of CYP4F3. These solutions are insensitive to changes in the inhibitor concentrations, which means that too high concentrations of the drugs do not cause negative side effects, which need to be considered in the administration of the final drug. A prioritisation of the targets in the LTB₄ branch, however, cannot be given as the necessary effective concentrations varied heavily for the different parametrisations of the model. Of these four targets, three have already been proposed as potentially relevant in the context of inflammatory diseases (see Table 1.3). Selective inhibitors of PGES seem to relieve pain [Xu et al., 2008], while inhibitors of LTA4H and 5-LOX seem to have an effect on cancer [Ding et al., 1999, Jeong et al., 2009]. This agreement of prediction and experimental results underlines the validity of my method for identifying potential drug targets.

Further treatments have been introduced in the results section but all of them seem less appropriate because they were heavily sensitive to changes in the concentrations of some of their drugs. In the context of inhibitors of the cyclooxygenases this seems quite restrictive as e.g. the treatment with aspirin would not have been selected as being acceptable. This might have two different reasons. First, the model might not be perfect and might require improvement. Second, the objective of keeping the PGI₂-TXA₂ ratio within 20% of its original value might be too restrictive to allow for biologically relevant treatments to be selected. Therefore, further simulations and experiments including COX inhibitors could lead to a clarification of this point. This is however beyond the scope of this thesis.

Points of improvement on the model The model of Yang *et al.* is not only suitable for drug target identification, but it also shows how drug development, Systems Biology, and experiments can and should go hand in hand. Predictions made in the results section of this chapter yet do not have the same quantitative quality as in the application of the sleeping sickness because of the parameter uncertainty associated with the model. The fact whether a treatment works better than another or whether it works at all is highly dependent on the parametrisation of the model. A possible next step in a cycle of predictions and experiments would therefore be to experimentally

test different combinations of existing inhibitors and use these results for the refinement of the model. Through further results it might be possible to reject some of the proposed parametrisations, which would it turn allow for a more quantitative assessment of the quality of different treatments.

Apart from using drug target predictions to improve the model, other problems of the model need to be tackled in the future. A very obvious one is the inconsistency between different descriptions of the model, i.e. the published SBML models and the description in the Supplement of Yang *et al.* . The most important inconsistency is the initial concentration of arachidonic acid in platelets. Both descriptions differ in three orders of magnitude. Calculations performed in this thesis are based on the high concentration, which is present in the model code. This high concentration is an important factor contributing to the fact that an inhibitor of the phospholipase alone cannot fulfill the objective because it cannot stop the TXA₂ production in platelets. Changing in the model at this point could therefore have a strong effect on the inhibitor prioritisation.

Furthermore, the model could be extended into various directions. Leukotrienes and prostaglandines could be exchanged between different cell types as it is a known fact that some cell types release eicosanoids to feed other cells that do not contain the corresponding enzymes to produce them [Folco and Murphy, 2006]. Another possible extension is the inclusion of the epoxygenase pathway, whose metabolites stem from AA and are also supposed to play a role in the inflammatory response [Spector, 2009].

5.4.4 Drug resistance through mutations

Depending on the kind of disease which is meant to be cured with a certain drug, resistance might become an important factor in the treatment's success. During the treatment of e.g. cancer or bacterial infections mutations might arise in the population of treated cells, which render them less sensitive to the applied drugs. This might be due to various facts: A mutation can cause a conformational change in a drug's target, reducing the probability of binding, the pathway being targeted can be made obsolete by a different, compensating pathway, which is activated, or the uptake of the drug into the cell can be reduced. An important question in this context is whether and how our choice of a proper target influences the development of resistances. Once this connection is made, targets can be specifically selected for lowering the probability of development of treatment insensitivities.

Based on a simple idea that is explained in section B.2.5 in the Appendix, it is possible to relate the necessary effective inhibitor concentrations, which are required to efficiently inhibit a target, to the probability of developing a

5.4. DISCUSSION

resistance against the drug. Although this is a heavily simplified model, it allows to draw the intuitive connection that targets, which are essential to the system and require very little inhibitor, need more mutations to become insensitive to the treatment. Therefore, resistances are less likely to develop in this case and we are given a second reason to prefer drug targets requiring lower effective inhibitor concentrations.

5.4.5 Outlook

5.4.5.1 Implementation

The identification of drug targets involves performing dynamic simulations of ODE models with many different sets of parameter values. Therefore, being able to rapidly calculate time courses for these models renders mandatory. Discussions with colleagues have taught me that the compilation of a model as a driver for an ODE solver is the best approach to tackle this problem. Following this idea it is not possible to make use of many of the available simulation tools as they only allow for dynamic simulations of single models or complex tasks like parameter estimations using a restricted set of algorithms. The employment of ODE solver libraries like LSODA within a custom tool, however, has provided reliable and sufficiently fast results.

A further improvement on the computational speed of my tool, can possibly be provided by interfacing the ODE solver library SUNDIALS [Hindmarsh et al., 2005] through the SBMLodeSolver [Machné et al., 2006], which, according to the author of the tool, is currently supposed to be the fastest solution to repeatedly solve ODE systems on a CPU [Machné, 2012].

5.4.5.2 Tackling parameter uncertainties

When modelling the dynamic behaviour of a biological system, one is facing different kinds of uncertainties. It can be questioned, whether the model comprises all relevant physiological processes, whether it describes molecular mechanisms correctly, and whether it does so with reasonable kinetics and accurate parameter values. Although research on metabolic pathways has almost ruled out structural uncertainties on the underlying reaction networks, experimentally determined kinetics and their parameter values may still contain measurement errors. Furthermore, differences in measurement conditions lead to a high variance in values reported in the literature.

These uncertainties on the parameter level can in principle be dealt with by performing a desired analysis over and over again for different parameter sets [Achcar et al., 2012]. Following such a procedure, one will end

up with an ensemble of results, which can be statistically analysed [Murabito et al., 2011]. Therefore, a potential extension of my framework, which deals with parameter uncertainties, could be implemented with the following steps. First, thermodynamically sound parameter sets could be drawn from biologically relevant distributions [Lubitz et al., 2010], second, the drug target identification can be repeated for each parameter set, and finally, a statistical analysis of the quality of the drug targets can be presented to the user. This statistical analysis could involve mean and variance of the necessary effective inhibitor concentrations as well as information on the fact how often such an inhibitor concentration could be found.

5.4.5.3 Using non-identifiabilities to direct further research

So far, non-identifiabilities of drug targets have only been considered to provide alternatives for targets against which no drug can be designed or which have to be neglected because of other reasons. Apart from this use identifiability analyses can further provide hints to guide research and should therefore be considered to be a part of the cycle of Systems Biology [Raue et al., 2009].

General non-identifiabilities in parameter estimations can have different causes. First, the model structure can comprise too much detail and can incorporate too many parameters to allow for a reliable estimation of the parameter values from measurable experimental data. Second, experimental data can be too sparse and may therefore not allow to discriminate between different parameter sets. Finally, also the values of parameters which are not estimated can play a role because they might force the trajectory of the system into certain areas in the state space which show a low dynamic complexity and therefore do not allow for parameter discriminations.

In the application of drug target identification, the reasons for non-identifiabilities are similar, yet slightly different. On the one hand, the considered pathway might be complex and allow for different inhibitions resulting in the same desired effect. On the other hand, the amount of experimental data being considered might be sparse. This translates into the question whether enough information on the discrimination between the “healthy” and the “diseased” state of the system is taken into account. Knowledge on this lack of information might then drive further research: Non-identifiabilities between several drug targets can be used to find differences in the dynamic response to drugs against each target. Then, information on differences in dynamic changes of certain observables can be used to direct new experiments. And finally, results of these experiments enter the objective function and potentially resolve the non-identifiability. With this cycle of experiments

5.4. DISCUSSION

and numerical simulations the predictive capabilities of the drug target identification will increase and eventually lead to the identification of potent drug targets.

5.4.5.4 Reverse drug target prediction enables mode-of-action identification

The applications of my framework presented in this thesis focus on the biological objective of identifying potential drug targets, i.e. enzymes whose inhibition leads to a certain desired effect, the transition from the “diseased” to the “healthy” state. Another possible application of the framework is the prediction of the mode-of-action of a drug. Such predictions can be valuable for drugs which have been identified by high-throughput screens and have therefore only been proven to work *in vivo* [Iorio et al., 2010]. Given a model and relevant experimental data of the treated organism before and after the drug has been applied, these two conditions can be defined as the “diseased” and the “healthy” state. The target identification applied to this data would then lead to targets and inhibition mechanisms which are able to explain the experimentally observed changes best and thus predict the drug’s most probable mode-of-action.

Detailed acknowledgements

I would like to thank Barbara Bakker for general supervision, providing the code for the Albert model, and for sharing unpublished work on the theoretical basis and practical applications of the network selectivity. Furthermore, I would like to thank Fiona Achcar for sharing a refined version of her model and Rainer Machné for discussions on implementation details of interfaces to ODE solvers.

Chapter 6

Drug-drug interactions

Contents

| | | |
|------------|------------------------------------------------------------|------------|
| 6.1 | Introduction | 165 |
| 6.1.1 | Synergisms and antagonisms in current research | 165 |
| 6.1.2 | Mathematical definitions of synergy | 168 |
| 6.2 | Methods | 171 |
| 6.2.1 | Synergism detection in ODE models | 171 |
| 6.2.2 | Choice of the null models | 171 |
| 6.2.3 | Computational detection of synergisms | 172 |
| 6.3 | Results | 174 |
| 6.3.1 | Glycolysis in <i>Trypanosoma brucei</i> | 174 |
| 6.4 | Discussion | 177 |
| 6.4.1 | Synergisms in the trypanosomal glycolysis | 177 |
| 6.4.2 | Advantages of the employed method | 178 |
| 6.4.3 | Advantages of synergisms and antagonisms for drug research | 179 |

6.1 Introduction

6.1.1 Synergisms and antagonisms in current research

The results of the last chapter have suggested that the methods presented in this work are suitable to identify treatments which are effective and safe.

6.1. INTRODUCTION

Amongst equally effective solutions, I have so far prioritised treatments containing fewer drugs and requiring the lowest drug concentrations. Additionally, with the help of non-identifiability analysis one can determine drugs that can replace each other in a certain treatment, which is an important information when preferred enzymes cannot be targeted.

In contrast to prior considerations, I will throughout this chapter explicitly drop the idea of achieving an objective with the fewest number of drugs possible. Instead, I will investigate combinations of drugs, which could in principle have the same effect, but do achieve it better than the individual drugs. The question of what exactly can be regarded as better and why will be answered in the following paragraphs.

6.1.1.1 Synergistic drug combinations require lower doses

A synergism between two or more entities generally means that the whole is bigger than the sum of the individual parts. In the context of drugs, it means that a treatment involving two synergistic drugs is stronger than what was expected from experiments involving only one of them. Such a synergistic combination can be used for two different purposes. First, one can use both drugs to achieve an effect that the single drugs cannot produce. Second, one can reduce the amount of the individual drugs given with each dose, which might reduce side-effects.

With the help of such synergisms it might also be possible to overcome resistances, as it has been shown that cells which are insensitive to two individual drugs can still be susceptible to a treatment involving both of them [Di Gaetano et al., 2001]. This idea has already been picked up for systematic screenings. Spitzer *et al.* have screened a compound library for having a highly synergistic action with fluconazole in different fungi in order to deal with resistance against this compound [Spitzer et al., 2011].

Another advantage of synergistic drug combinations is that they are likely to be organism-specific, which has been shown experimentally for synthetic lethal interactions [Tischler et al., 2008]. From a theoretical basis this can be explained by the fact that synergisms seem to be highly dependent on the network's structure [Lehár et al., 2007]. Numerical parameter values have a less prominent effect on synergisms except if they significantly change the model's behaviour [Fitzgerald et al., 2006, Jackson, 1993]. Therefore, knowledge about combined chemical effects and the targets of those chemicals can even be used to predict an unknown network structure [Segre et al., 2004, Lehár et al., 2007, Nelander et al., 2008].

6.1.1.2 Antagonisms and resistance

In contrast to synergy, an antagonism between two or more entities describes the circumstance that their combination is worse than what was expected from the single parts. For drug combinations strong antagonisms can even mean that the combination of two drugs has a smaller effect than each drug individually. The fact that such a combination of drugs can be interesting is opposing all previously made assumptions on preferable treatments. However, they have a very important application.

Treatments targeting cells that are rapidly evolving, e.g. microorganisms or cancer cells, are often facing the problem of drug resistance development. It has been determined experimentally and theoretically that antagonisms between drugs in a treatment can delay resistance development [Chait et al., 2007, Michel et al., 2008, Hegreness et al., 2008, Yeh et al., 2009]. If a treatment with two drugs leads to less severe consequences than treatments with the individual drugs, then resistance to one of the drugs under the combination treatment will render an individual less fit than the wild type, thus delaying complete resistance development. Thus, strongly antagonistic drug combinations might be advantageous for the treatment of certain diseases.

6.1.1.3 What causes synergy?

Synergisms and antagonisms which might be desirable for the aforementioned reasons can have various causes. First, they can target the same protein such as gefitinib and certuximab. These drugs are binding the EGF receptor ErbB1 in two different sites [Matar et al., 2004] and provide a prolonged inhibition of the receptor in the treatment of cancer. Second, they might target different proteins involved in the same or a functionally related pathway. For diclofenac and paracetamol, as for many other combinations of painkillers, it has been supposed that they inhibit different isoenzymes, but their synergistic effect could also be caused through individual effects in various off-pathways [Miranda et al., 2006]. A third cause of synergy can appear on the population level. Here, individual drugs might attack different subpopulations which have already acquired partial resistances, thereby eradicating the population completely. An example for this kind of synergy is the action of cisplatin and mitomycin C in the treatment of cancer [Durand, 1989]. A final cause of synergy can occur on the pharmacokinetic level. One drug can increase the uptake and the distribution of another drug, it can alter how the drug is metabolised, and delay its excretion. All of these processes increase the bioavailability of the active substance, which leads to a stronger and prolonged effect as it is the case for cyclosporin A and paclitaxel [Bardelmeijer

6.1. INTRODUCTION

et al., 2000].

For my predictions on the quality of drug targets, I have so far used only models of single pathways. Thus, it will be outside the scope of this work to identify synergies other than those of the second type. However, it should be noted that most synergistic drug pairs identified in high throughput screens are more likely to be synergistic because of side-effects [Cokol et al., 2011]. Considering this, one cannot expect to be able to find all possible drug synergisms on the basis of ODE models of single pathways.

6.1.2 Mathematical definitions of synergy

6.1.2.1 Null models of combined effects

If one defines drug synergisms and antagonisms as combinations performing better or worse than expected from the individual effects, then one first has to define what actually is expected. According to Chou more than 300 equations have been published that try to quantify drug combination effects in various situations [Chou, 2010]. A few simple and accepted null models of how a neutral drug interaction should look like are introduced in the following. The interested reader is referred to [Greco et al., 1995] for more details on these and further information on other models.

Highest single agent This model assumes that whenever two drugs are applied in parallel, their combined effect is the maximum of their single effects. Such an understanding of a drug interaction can be related to the idea of a “bottleneck” in the flux of information or matter through a system. If two drugs are used in parallel, one of them will represent the more narrow bottleneck and will therefore determine the effect of the treatment.

Loewe additivity Loewe and Muischnek have come up with a model of neutral drug interaction which they termed additivity [Loewe and Muischnek, 1926]. This model assumes no interaction between the drugs and is best described by the fact that a drug being tested for synergy with itself would be considered additive. What it basically assumes is that if one achieves a certain effect with a single drug at a certain concentration, one will achieve the same effect if a fraction of the first drug is replaced by the same amount of the second drug. In mathematical terms two drugs are additive if they fulfil the equation

$$1 = \frac{C_{A,x}}{IC_{A,x}} + \frac{C_{B,x}}{IC_{B,x}}, \quad (6.1)$$

6.1. INTRODUCTION

where $IC_{A,x}$ is the individual dose of A needed to achieve effect x and $C_{A,x}$ is the dose in combination with drug B to achieve the same effect. Despite the age of this model, so far there has been no substantial criticism to it [Greco et al., 1995].

Bliss independence Bliss assumes in his null model, which is termed independence, that fractional effects of two drugs multiply to give the combined effect [Bliss, 1939]. This means that if one drug reduces the production of an observed output to half of its normal value and another drug to one third, the observable's production should decline to one sixth under a treatment involving both drugs.

In contrast to the Loewe additivity, people have criticised Bliss independence. One argument is that the application of one drug will most probably alter the dynamic behaviour of the system. Therefore, it cannot be assumed that the system's response to the second drug will be the same when it is applied in combination with the first drug or without [Gessner, 1974]. However, I would argue that this behaviour is actually an interesting point to detect and that I would like to use synergism analysis to detect cases in which one drug provides a second drug with an "environment" in which it can work more effectively. A second argument against this model is that the same drug tested for synergy with itself is not regarded as neutral (as it would be with the Loewe additivity) [Grindey et al., 1975]. This point is true, and one should therefore not investigate drugs targeting the same enzyme for synergy using Bliss independence.

6.1.2.2 How to detect synergy

Given the aforementioned null models of drug interaction, it is possible to define synergistic and antagonistic drug combinations as treatments being more or less effective than expected. However, depending on which null model is used, the ways of how synergisms are detected do differ.

Isobolographic analysis The idea of isobolographic analysis is very simple in its theory. In a plot where the two axes represent the concentrations of the compared drugs, one draws a line through points representing different concentration combinations leading to the same effect. If this line is straight, the drugs follow the Loewe additivity, if the line is bent towards the origin of the plot, they are synergistic, and if it is bent away from the origin, they are antagonistic [Loewe and Muischnek, 1926].

6.1. INTRODUCTION

Fractional product method Compared to the isobolographic analysis, which asks for drug concentrations leading to the same effect, the fractional product method [Webb, 1963] compares different effects for the same drug concentrations. Given fixed drug concentrations for the two drugs, the method predicts their combined effect by their individual effects using the formula

$$fu_{12} = fu_1 \cdot fu_2, \quad (6.2)$$

where fu is the fraction unaffected of either combined or individual treatment. This fraction unaffected represents the variable that should be reduced by the treatment, e.g. the production of a metabolite in a certain pathway compared to the untreated state. If the effect of the combined treatment is stronger than expected from the Bliss independence, the combination is synergistic, if it is weaker, it is antagonistic. This decision can in principle be statistically assessed for the confidence associated with it [Drewinko et al., 1976, Prichard and Shipman, 1990], however, I will not go into detail on this point.

6.1.2.3 How to quantify synergy

For the application of rating drug combinations, one does not only need to know whether they are synergistic or not, but one needs to quantify this synergy. Depending on the chosen null model, this quantification is represented by different formulas.

For the Loewe additivity, a quantification can be given through the interaction index [Berenbaum, 1977]. The formula

$$I = \frac{C_{A,x}}{IC_{A,x}} + \frac{C_{B,x}}{IC_{B,x}} \quad (6.3)$$

is an extension of Equation 6.1 and quantifies how far the same-effect-isobole is off from being a straight line at a certain angle. If the interaction index is in the range $1 < I < \infty$, higher values signify stronger antagonisms, and if it is in the range $0 < I < 1$, lower values indicate stronger synergy. This principle has been taken up and extended many times in the literature. The most prominent of these extensions is the median effect analysis [Chou and Talalay, 1984], which extends the null model and describes how synergies can be predicted from little experimental data.

For the Bliss independence, a quantification can be given by the Bliss boost model (e.g. [Lehár et al., 2007, Yeh and Kishony, 2007]). This model

extends Equation 6.2 by rewriting it in terms of the affected fractions ($fa_x = (1 - fu_x)$)

$$fa_{12} = fa_1 + fa_2 - fa_1 \cdot fa_2$$

and then adding a factor α , which represents the degree of cooperativity

$$fa_{12} = fa_1 + fa_2 - \alpha \cdot fa_1 \cdot fa_2. \quad (6.4)$$

If the cooperativity is in the range $-\infty < \alpha < 1$, lower values represent stronger synergy, while in the range $1 < \alpha < \infty$, higher values indicate stronger antagonisms.

Apart from these two basic approaches, lots of other quantifications of synergy exists in the literature (see [Greco et al., 1995] for a review) and even combinations of the aforementioned approaches are possible (e.g. [Berenbaum, 1985]).

6.2 Methods

6.2.1 Synergism detection in ODE models

Synergisms in ODE models have already been investigated in various cases with diverse approaches [Jackson, 1993, Fitzgerald et al., 2006, Lehár et al., 2007]. However, to my knowledge no tool being capable of investigating them in general reaction networks is freely available. In this chapter I will present the methods which are the fundamental core of the numerically stable synergism detection implemented in my tool. These methods are largely based on the ideas presented by Lehár *et al.* [Lehár et al., 2007], they are, however, adapted to the use in an automatic tool.

6.2.2 Choice of the null models

Depending on which kind of experimental information they require, the null models presented in the introduction of this chapter fall into two different groups. The first group is derived from the Loewe additivity and in order to check for synergisms or antagonisms in a drug combination one is asking for different drug concentrations leading to the same effect. The other group is derived from the Bliss independence and utilises the effects of combinations of drugs with constant concentrations to test for synergy.

Thus, depending on which class of null model one wants to investigate different simulations have to be carried out. For the Bliss-based models,

6.2. METHODS

chequerboard analysis can be performed, which is the extension of the fractional product method to a whole grid of drug concentration combinations. Following this approach has the advantage that all necessary simulations are known in advance and can be performed in parallel on a cluster. On the contrary, the Loewe-based models require the search for drug concentrations having a certain effect. This is essentially a parameter estimation problem and might therefore require a much larger number of sequential simulations, which makes the chequerboard approach more attractive from the computational side.

In principle, none of the generally accepted models has any kind of justification on why it should universally be able to predict the effect of drug combinations in complex reaction networks [Greco et al., 1995]. However, Lehár *et al.* have been able to show that certain structural patterns in reaction network can lead to combination effects that can be explained by Bliss-based null models. Therefore, I will mainly use this kind of models in the following.

6.2.3 Computational detection of synergisms

Chequerboard analysis For my analyses on the synergistic action of drug pairs, simulations are carried out on a grid of drug concentrations for each available drug pair. The grid consists of ten concentration points for every drug: 0 and nine logarithmic steps centred around a relevant concentration [Borisov et al., 2003]. As a relevant concentration, I try to estimate the drug's EC50 value, which is the inhibitor concentration at which 50% of its maximal effect is achieved. This concentration has been chosen because around it the effects of the individual drugs change most significantly and, therefore, it is most informative. Furthermore, in order to avoid numerical problems, extreme values for the relevant concentration, i.e. below 0.1 or above 1000, are avoided and the relevant concentration is in these cases set to 10.

Synergism model fitting Given the simulation results of the chequerboard analysis and the parametrised synergism models, one can optimise parameters of these models to make their predictions for a combined effect fit the simulated data. For most of the models an optimisation step is not required as they have no parameters. These models include single agent models, describing that only one of the drugs is responsible for the combined effect, highest single agent, and the Bliss independence. For parametrised models as Bliss boost and optimisation of the cooperativity factor is carried out using the differential evolution heuristic [Storn and Price, 1997].

Integrating Loewe additivity Using the results of the chequerboard analysis, tests for Loewe additivity are not easily possible. As an approximation of the computations needed to detect Loewe synergisms, I perform the following calculations, which are based on ideas presented in [Berenbaum, 1985]. I first predict the combined effect of the combination treatment by the effect of one of these drugs in the combined concentration. If the effect at this certain concentration is not provided on the chequerboard, the value is linearly extrapolated in a second step from the closest values available. Using this method, Loewe additivity can approximately be tested for using chequerboard simulations.

Model selection Finally, my method wants to be able to judge, whether a combination of drugs is synergistic or not. In order to do so, the following steps are performed. First, the model best describing the combined effects has to be found. If we compare models without parameters, this can easily be done by taking the model with the best goodness of fit. However, in order to be able to compare models with different numbers of parameters, my tool displays the Akaike information criterion (AIC) [Akaike, 1974] for each model fit. This criterion signifies whether a more complex model, i.e. one including more parameters, is really necessary to describe the data. Given that the model best describing the data contains parameters, these can finally be used to quantify the degree of synergism or antagonism in between two drugs.

How to quantify a drug's effect In the previous chapter a drug's effect has always been measured in terms of the objective function, the residual sum of squares between desired concentrations at certain time points and their simulated concentrations. The synergism models in this chapter, however, measure the effect in a linear variable, e.g. as the fraction of a population alive after treatment or the relative production of a certain toxic substance via a pathway. Therefore, the cooperativity models cannot be easily transferred to the framework described beforehand. A more detailed analysis of this fact is shown in section B.3.1 in the Appendix for the Bliss boost model. There it is also shown that the application of the Bliss boost model to the residual sum of squares is practically only possible if the objective function is made from a single time point. For all other purposes, the drug effect has to be measured by a linear function.

6.3. RESULTS

Table 6.1: Synergisms between non-competitive inhibitors in the trypanosomal glycolysis. The results show all combinations for which the Bliss boost model fitted the data best (according to the Akaike Information Criterion (AIC)), which acted significantly synergistically (cooperativity < 0), and which had a relevant effect on the flux through the glycolysis (objective function could be reduced to less than 1, i.e. more than 50% flux reduction).

| Drug 1 | Drug 2 | AIC | Model | Cooperativity | Min. obj. val. |
|--------|--------|------|-------------|------------------------|----------------------|
| nPFK | nGAPDH | 3020 | Bliss boost | $-1.22 \cdot 10^{-8}$ | $1.93 \cdot 10^{-8}$ |
| nPFK | nALD | 3050 | Bliss boost | $-9.17 \cdot 10^{-9}$ | 0.0924 |
| nPFK | nGPDH | 3050 | Bliss boost | $-2.71 \cdot 10^{-9}$ | 0.0914 |
| nPFK | nPGK | 3050 | Bliss boost | $-1.88 \cdot 10^{-9}$ | 0.0183 |
| nPFK | nPT | -509 | Bliss boost | $-2.00 \cdot 10^{-11}$ | 0.106 |
| nPFK | nPGM | -520 | Bliss boost | $-2.00 \cdot 10^{-11}$ | 0.0984 |
| nPFK | nENO | -531 | Bliss boost | $-2.00 \cdot 10^{-11}$ | 0.0726 |

6.3 Results

In order to test the validity of the computational method presented in this chapter the artificial model presented in [Lehár et al., 2007] has been investigated. The method has been able to reproduce the results of Lehár *et al.* with respect to the synergism models implemented in my tool. These results implicate a high dependency in between the structure of the underlying reaction network and the way in which drugs act in combination on this pathway. However, they are not shown in this work as they do not provide novel information.

6.3.1 Glycolysis in *Trypanosoma brucei*

As a real biological application I investigate synergisms and antagonisms in the trypanosomal glycolysis as implemented in the Albert model [Albert et al., 2005]. A visualisation of this model is given in Figure C.6 in the Appendix. Using the synergism detection heuristic described in the methods part of this chapter I investigate the combined effects of different non-competitive inhibitors on this network. For all further considerations involving this model the effect is quantified by the square root of the objective function as presented in Equation C.4 in the Appendix.

6.3.1.1 Detected synergisms

Table 6.1 shows the results of the synergism analysis for the trypanosomal glycolysis. From the result list drug combinations following an interaction other than the Bliss boost model and combinations without a significant

Table 6.2: Antagonisms between non-competitive inhibitors in the trypanosomal glycolysis. Results show again combinations for which the Bliss boost model fitted best, which show a significant antagonism (cooperativity > 2), and lead to a relevant reduction of the pathway's activity.

| Drug 1 | Drug 2 | AIC | Model | Cooperativity | Min. obj. val. |
|--------|--------|------|-------------|---------------|-----------------------|
| nPFK | nAKc | 3140 | Bliss boost | 100 | $4.19 \cdot 10^{-10}$ |
| nPFK | nAKg | 3020 | Bliss boost | 100 | $5.67 \cdot 10^{-8}$ |
| nGK | nGPO | -321 | Bliss boost | 37.4 | 0.0879 |

effect on the objective have been removed. As the Bliss independence model seems to judge the combined effects of two drugs on this network better than the Loewe additivity, only Bliss synergisms are regarded in this context. Among the resulting 7 drug combinations, two subgroups with different AIC values showed up in the results. The high AIC values of the first results indicate that the Bliss model could not be fitted accurately to the data. A closer inspection of the simulation results show outliers in the simulated treatments, which indicate numerical problems of the solver. Such numerical problems have already been encountered beforehand with inhibitors of PFK and are not a problem of the synergism detection.

For the remaining three results the fit to the Bliss boost model is significantly better. However, also among these combinations simulations have run into numerical problems with the non-competitive inhibitor of PFK. Whenever the concentrations of this inhibitor is raised above 15, simulations crash and lead to outliers in the results matrix. These outliers severely affect the synergism model fitting, as a significant part of the results matrix is incorrect. If one disregards these outliers, the HSA model would best describe the interaction of the three drug combinations.

Thus, according to my prior criteria no synergism being in agreement with the Bliss boost model could be identified in the Albert model [Albert et al., 2005].

6.3.1.2 Detected antagonisms

Table 6.2 shows the results of the antagonism identification. These results again contain two subgroups, one with high AIC values, which again stem from failed simulations, and the combination of non-competitive inhibitors targeting glycerol kinase (GK) and glycerol phosphate oxidase (GPO). This combination is comparably well fit by the Bliss boost model and does not contain an inhibitor of PFK, whose simulation is particularly prone to numerical errors. Therefore, it is investigated in more detail.

A closer look into the way how different drug interaction models fit the

6.3. RESULTS

Table 6.3: Fitting results of different synergism models to the simulation data from simultaneous inhibition of GPO and GK. The different versions of the Loewe additivity model stem from the fact to which single inhibitor the combined drug effect is compared.

| Model | AIC | Sum of residuals | Sum of squared residuals | Model variables |
|--------------|-------|------------------|--------------------------|-----------------|
| Bliss boost | -321 | -1.31 | 0.411 | $\alpha = 37.4$ |
| HSA | -66.3 | 29.1 | 22.7 | |
| Second alone | -66.3 | 29.1 | 22.7 | |
| Bliss | -61.0 | 32.1 | 24.7 | |
| Loewe Y | -37.9 | -2.76 | 35.4 | |
| First alone | 32.2 | 70.9 | 106 | |
| Loewe X | 32.9 | 71.4 | 107 | |

simulated data of GK and GPO reveals that the interaction is best described by a strong Bliss boost antagonism (Table 6.3). Judging from the AIC, this model fits the interaction by far better than the other models.

Detailed simulation results of the drug combination are given in Figure 6.1 together with the interaction prediction of the Bliss boost model and estimates of the cooperativity in the space of drug concentrations. These results point to two interesting facts. First, the drugs show a stronger combined effect for concentrations over 10 when compared to the effects of the single drugs. The fact that this cooperativity is regarded as an antagonism is driven by the response of the pathway to non-competitive inhibitors of GK, which are increasing its activity. According to Equation 6.4, the combination is therefore classified as being antagonistic although it has a strongly negative impact on the flux through the pathway. Second, although the Bliss boost model fits the data comparably well, the model is partially over- or underestimating the antagonistic effect depending on the concentration of the GPO inhibitor. Because the cooperativity only seems to take effect after the inhibitor has risen above a concentration of around 1, lower concentrations show no antagonistic behaviour and the cooperativity matrix seems to be divided into two distinct parts. This indicates that a model being sigmoidally dependent on the GPO inhibitor concentration might be more suitable to describe the combined effect of these drugs.

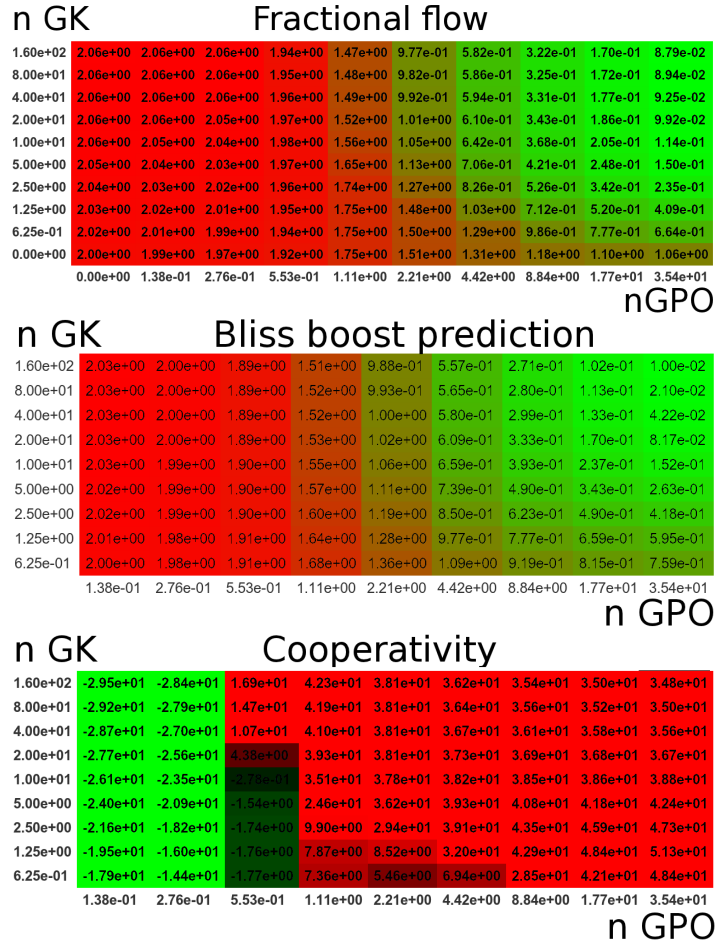


Figure 6.1: Results of the synergism analysis in between non-competitive inhibitors of GPO and GK. The heat maps show the simulated effect of the inhibitors on the objective function for varying inhibitor concentrations, the predicted effects according to the fitted Bliss boost model, and the cooperativity factor of the Bliss boost model if all data points were fitted individually to the Bliss boost model.

6.4 Discussion

6.4.1 Synergisms in the trypanosomal glycolysis

In conclusion, inhibitors of GPO and GK are the only drug combination showing a relevant cooperativity. Parallel inhibition of the glycerol phosphate oxidation and the glycerol kinase does not only lead to a complete inhibition of the glycolysis, but it is achieving this objective using relatively

6.4. DISCUSSION

low inhibitor concentrations. In fact, this combined effect is already known in the literature and can be exploited using for example salicylhydroxamic acid (SHAM) together with glycerol [Clarkson et al., 1981]. SHAM is an inhibitor of the trypanosomal alternative oxidase (TAO), which is an enzyme necessary for the oxidation of glycerol phosphate, while glycerol, the substrate of the glycerol kinase, is inhibiting the flux through glycerol kinase thermodynamically [Krakow and Wang, 1990].

The mechanism by which the synergistic combination of a GPO and a GK inhibitor work is fairly complex when compared to the single targets in the trypanosomal glycolysis. All of the single targets work by reducing the capacity of the main route of carbon flux from external glucose to cytosolic pyruvate. Both enzymes, GPO and GK, are not part of this route and therefore act differently. Along the carbon flux route, the cofactor NADH is produced by GAPDH. If the organism is unable to convert NADH to NAD⁺ in the glycosome, the reaction catalysed by GAPDH (and therefore the whole glycolysis) cannot proceed indefinitely. Under normal conditions this cofactor conversion is done by the GPDH, which in turn produces glycerol 3 phosphate. This product is either consumed by the GK or it is converted back to DHAP, which involves TAO. Thus an inhibition of these two enzymes leads to an increase in glycerol 3 phosphate making the reaction catalysed by GPDH thermodynamically unfavourable and depleting NAD⁺, without which the glycolysis cannot proceed.

6.4.2 Advantages of the employed method

The method described in this chapter is based on work of Lehár *et al.* [Lehár et al., 2007] and is able to automatically search for potential synergisms and antagonisms in between hypothetical drugs in a given model. It is built on accepted models of drug interactions and is able to evaluate Loewe-based and Bliss-based types of synergism models. When compared to other steps proposed in this thesis, synergism analysis does only require little more computational effort than 2d inhibitor scans. It requires the determination of EC50 values and the fitting of interaction models to simulation results, which are in principle small parameter estimation problems with few parameters.

In spite of the inclusion of many different steps and methods into my synergism detection heuristic, it is unable to provide spotless results without user interaction. The results part of this chapter has shown, that computational problems along the simulations of inhibitions might occur. Those numerical problems can affect the results and lead to a corrupted synergism prediction. A further problem is the question of how to automatically judge the goodness of fit in between an interaction model and simulation results.

The AIC provides means of how different models having varying parameter numbers can be compared, however, it can not be used to tell how good a single model fits the simulation results.

Compared to the very simple synergism prediction I have published beforehand [Schulz et al., 2009], criteria on what should be regarded a relevant synergism have become more strict. The results from this article included the GPO/GK pair but also further pairs, which have not been able to achieve a therapeutically relevant reduction of the objective function. This explains why the current list of synergisms and antagonisms is much smaller than my previously published one.

6.4.3 Advantages of synergisms and antagonisms for drug research

Knowledge on synergistic or antagonistic actions of drugs can provide valuable information for the development of successful treatments. Synergisms, which achieve a certain effect with multiple inhibitors in low doses, can be either used to produce stronger or prolonged responses to a treatment or to reduce side-effects caused by a high concentration of one drug. In contrast, antagonisms require higher doses to achieve an effect but antagonistic drug combinations can delay the development of resistances against a treatment.

As the computational effort of detecting synergisms and antagonisms does only slightly increase compared to the two-dimensional inhibitor scans, this type of analysis should be performed whenever possible. It can be applied to a drug target identification problem, if the objective either follows the restrictions mentioned in section B.3.1 in the Appendix or if it can be transformed into a linear function. A final remaining problem, however, is the decision on whether a synergism or an antagonism is favourable for a treatment. This decision surely depends on the specific disease which is to be treated and should be decided individually for every case.

6.4. DISCUSSION

Chapter 7

Discussion

Contents

| | |
|------------------------------------------------------------|------------|
| 7.1 Achievements | 181 |
| 7.1.1 A framework for drug target identification | 181 |
| 7.1.2 Target predictions | 183 |
| 7.2 Directing new research | 184 |
| 7.2.1 How to proceed from predictions | 184 |
| 7.2.2 Curing sleeping sickness | 184 |
| 7.2.3 Reducing inflammatory responses | 185 |
| 7.3 Extensions to my framework | 185 |
| 7.3.1 Using web resources for drug prediction | 185 |
| 7.3.2 Considering parameter uncertainties | 187 |
| 7.4 Systems Biology for drug research | 188 |
| 7.4.1 Comparing different approaches | 188 |
| 7.4.2 Limitations of predicted drugs | 189 |

7.1 Achievements

7.1.1 A framework for drug target identification

Within this thesis I have introduced a workflow which aims at the unification of further efforts using ODE models of biochemical processes as an aid in the search for efficient and safe drug targets. This workflow, which is visualised in Figure 7.1, consists of five building blocks, three of which

7.1. ACHIEVEMENTS

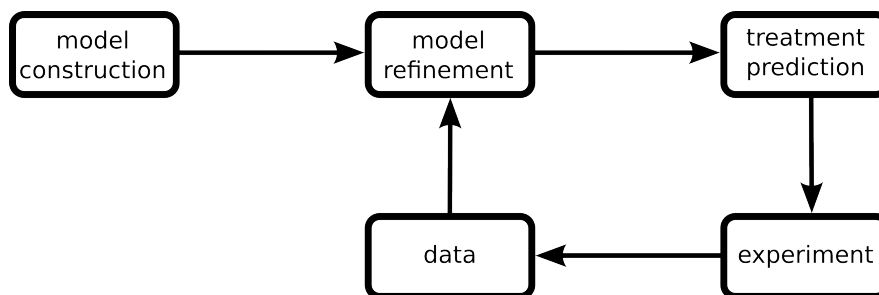


Figure 7.1: Overview on the workflow of TIde for the identification of potent targets with the help of mathematical models.

involve computational steps that are enabled or simplified through methods and software introduced here. First, a model either needs to be constructed from scratch or taken from the literature. Second, an initial version of the considered model is refined and extended using available experimental data and computational models. Third, the model is investigated for potential drug targets, which lead to the reestablishment of “healthy” conditions upon inhibition or activation. Finally, these predictions are supposed to be tested in experiments, e.g. via known small chemical entities interacting with the targets or via RNA interference [Fire et al., 1998], a method for the specific reduction of single protein concentrations through the introduction of small, double-stranded RNA fragments [Ngô et al., 1998, Elbashir et al., 2001, Novina et al., 2004]. In accordance with the cycle of Systems Biology, the results of these experiments are then checked for being in agreement with the model. If the resulting data agrees with the model, further predictions can be made based on the current model and drugs can be developed against the selected targets. If the data contradicts the model, it has to be refined. This includes changes in parameter values, addition or removal of regulatory mechanisms, or alterations in the stoichiometry of the reaction network in order to make the model fit this new data.

The methods introduced in this work are designed to simplify the first three steps of this workflow: the search for relevant available models, the alignment and combination of models, and the detection of drug targets in them. Through the methods introduced in chapter 3 the software semantic-SBML provides means to find relevant models and experimental data starting from an initial model or data set. To my knowledge no other software is available at the time of writing which is able to achieve a similar objective. The most closely related methods are concerned with query based retrieval of computational models, i.e. a tool being able to search for models related to provided keywords [Henkel et al., 2010].

The aligning and combination of networks describing biochemical processes is not a novel idea in general. Our software semanticSBML, however, provides easily accessible yet methodologically advanced means to compare and combine reaction networks with the methods described in chapter 4. By treating the problem of model alignment in a way similar to the comparison of models, my methods lowers the computational burden of network alignment and allows for the structural comparison of moderately large models.

Finally, methods introduced in the chapters 5 and 6 have paved the way for the development of the software TId. To my knowledge this is the only open-source tool allowing for the identification of effective and safe drug treatments with the help of ODE models. Other published approaches do either rely on a different kind of model as an input (e.g. [Karp et al., 2010]) or do not publish their software (e.g. [Yang et al., 2008]). Furthermore, through the integration of synergism and non-identifiability analysis my tool provides a unique combination of established and novel methods, which have shown their applicability in the examples presented within this work.

7.1.2 Target predictions

Within this work I have applied my methods for the prediction of efficient and safe drug targets to examples which fall into two different categories. First, a number of results from the literature has been replicated to show the validity of my approach. These results include the investigation of synergistic drug actions in an artificial model and the potential drug actions in a simple linear pathway. Through the latter example I have been able to deduce general principles on the quality of different targets and different modes-of-action.

Second, I have investigated potential drug targets in established models of different pathways, which have already been analysed using other approaches. One of these examples is the glycolysis in *Trypanosoma brucei*. Various versions of this model have already been analysed and targets have been rated through metabolic control analysis. My results, which simulate the effects of realistic inhibitor concentrations on the model, do differ in small points from conclusions that have been drawn beforehand. Furthermore, they highlight parts of the models which are in need of more careful experimental validation and modelling. The other biologically relevant example provided throughout this thesis is the arachidonic acid pathway in different human cells. My results on this model agree with published data [Yang et al., 2008], however, they provide clearer insights into the way how successful treatments affect the dynamics of the pathway.

7.2 Directing new research

7.2.1 How to proceed from predictions

The cycle of Systems Biology implies that theoretical and experimental work always have to go hand in hand. As the main contributions of my work are of a purely theoretical nature, possible practical experiments implicated by my results should be discussed.

In general, as already implicated by the workflow in Figure 7.1, experiments are supposed to support or invalidate predictions of the drug target identification step. Therefore, the predicted treatments should be used to search the literature for existing inhibitors or activators against the proposed targets. If such inhibitors can be found their binding to the target protein needs to be quantified. Appropriate K_i values can either be found in public databases like KiBank [Zhang et al., 2004] or the literature, or they have to be determined experimentally. Furthermore, one should rule out the possibility that the applied inhibitors bind to multiple targets in the considered network or, if this is inevitable, account for this fact in further simulations. As an alternative to the application of small chemical entities to the system, which might not be available, RNA interference [Fire et al., 1998] can be considered. Reducing the amount of expressed protein through this technique can be considered to be equivalent to the treatment with non-competitive inhibitors, which effectively also reduce the maximal velocity of a reaction.

Subsequent to these considerations, the selected perturbations are applied in experiments. The experimental conditions under which the new results are gained should be planned such that they are closely related to the conditions under which other data that has been relevant for the construction of the model has been obtained. Finally, the resulting data is compared to model predictions. If the data agrees with the model, the successfully tested treatment could serve as an initial result for further drug development efforts refining drug molecule structures to optimise pharmaceutically relevant properties. If the data, however, does not agree with the model, this data can be used in the refinement of the model. This additional knowledge could point to the fact that some parts of the model, its structure, its kinetics, or its parameter values need to be changed to increase the model's predictive power.

7.2.2 Curing sleeping sickness

The results based on my analysis of the trypanosomal glycolysis in chapters 5 and 6 suggest a number of interesting drug targets that could be verified

7.3. EXTENSIONS TO MY FRAMEWORK

in experiments. First, the inhibitions of the glyceraldehyde 3-phosphate dehydrogenase, which are the most promising treatment, can be tested. This could be done by using RNAi to lower the protein concentration as done in [Albert et al., 2005]. Second, inhibitions of the trypanosomal hexose transporter have lost their influence on the pathway along newer versions of the glycolysis model. RNAi experiments could be used to clarify this influence and the resulting data might be an important step in the further refinement of the model. Third, simulations involving an inhibitor of the phosphofructokinase have lead to interesting results. Experiments involving the PFK inhibitor polycarpol [Ngantchou et al., 2009] could be performed and their results might lead to a further improvement of the model. Finally the proposed synergistic effect of SHAM and glycerol treatment could be quantified to further refine the oxygen consuming branch of the model.

7.2.3 Reducing inflammatory responses

Results on the arachidonic acid pathway provided in chapter 5 indicate that a number of experiments still needs to be done solely for the purpose of refining the current model. While it is clear from experiments and currently available treatments that a number of proteins are successfully targeted by different drugs, a prioritisation of these targets cannot be given by my analyses yet. The reason behind the “fuzziness” of my predictions is the fact that multiple parameter sets for the model have been published. As the quality of the different targets highly depends on these parameters, no reliable predictions can be made without further restrictions on their numerical values. The determination of further necessary experiments could be done using methods from optimal experimental design, e.g. [Vanlier et al., 2012], the application of such methods to the examples presented, however, is beyond the scope of this thesis.

7.3 Extensions to my framework

7.3.1 Using web resources for drug prediction

Years ago the identification of drugs against diseases has been solely in the hands of the pharma industry. The reason behind this gap between industry and academia has been the lack of public databases storing information on the bioactivity of large sets of compounds, which are required to move from knowledge about a potent target to a potent drug candidate. Over the last years this gap has been closed by publicly available databases, which now

7.3. EXTENSIONS TO MY FRAMEWORK

complement the knowledge available to academia [Bender, 2010].

7.3.1.1 Publicly available resources

Today, a diverse set of web resources containing information on drug-target relationships is publicly available. These knowledge bases differ in the information they are built upon and on the information and tools they offer.

A large fraction of these resources have been compiled using protein-ligand structures from the Protein Data Bank (PDB). Examples are Relibase¹ [Hendlich et al., 2003], which for example contains analyses of ligand similarities and allows for ligand substructure and interaction searches, and the Potential Drug Target Database² (PDTD) [Gao et al., 2008], which amongst other information contains binding affinities.

A second kind of web resources are built upon integrated information from commercial databases and literature information. Examples of such resources are the list of druggable protein domains [Hopkins and Groom, 2002, Russ and Lampel, 2005], DrugBank³ [Wishart et al., 2008], which apart from drug-target relations contains further pharmacologically relevant information like drug-drug interactions, or BindingDB⁴ [Liu et al., 2007], a database of experimentally determined drug affinities.

Due to the heavy increase in the number of available web resources containing information on drugs and their targets, meta-databases as STITCH⁵ [Kuhn et al., 2010b] have become available. These resources integrate interactions from metabolic pathways, chemical structures, or results of binding experiments into complex knowledge bases. In addition, different resources compiling indirect information on the action of drugs are available. An example of such a database is SIDER⁶ [Kuhn et al., 2010a], which collects drug side-effects compiled from package inserts.

7.3.1.2 Identifying drugs for selected targets

In succession to the prediction of potent drug targets using my or a different workflow, the gained target information can be enriched by potential inhibitors acting on the selected proteins. Knowledge about such drug candidates could then be used in upcoming experiments to support or falsify the predictions.

¹<http://relibase.ccdc.cam.ac.uk>

²<http://www.dddc.ac.cn/pdtd/>

³<http://www.drugbank.ca>

⁴<http://www.bindingdb.org>

⁵<http://stitch.embl.de>

⁶<http://sideeffects.embl.de>

7.3. EXTENSIONS TO MY FRAMEWORK

As the manual determination of potential drugs from the aforementioned web resources is tedious work when done thoroughly, this step should be supported by software. For a similar purpose, Cockell *et al.* [Cockell et al., 2010] have integrated diverse data sets from various sources into a comprehensive knowledge base. Then, they have constructed complex patterns of potential relations between entities of this new resource in order to determine potentially new applications for known drugs.

A similar idea could be used to find potential drugs for selected targets when combined with methods presented in this work. For this purpose I would propose to start from an annotated SBML model. In this model, kinetic and semantic information could be used to determine enzymes in the model structure automatically. Using the semantic similarity proposed in chapter 3 information on the function of an enzyme can be used identify the protein catalysing the reaction, which can in turn be used to identify drugs binding to it. After a successful target identification step, these results on potential modulators of the target enzymes can be presented to the user for further experimental validation.

7.3.2 Considering parameter uncertainties

As already mentioned in the discussion of the TIDE approach, model parameter can heavily affect the results of my methods. Therefore, uncertainties associated with them should explicitly be taken into account when targets are selected and prioritised. Computationally this can be done by sampling parameters from an assumed distribution and the distribution of the necessary effective inhibitor concentration of the targets could be analysed.

Furthermore, these parameter uncertainties could be used in the planning of new experiments. Optimal experimental design [Kreutz and Timmer, 2009, Vanlier et al., 2012] could be used to dedicatedly select those experiments that can reduce the uncertainty associated with the parameters having the largest impact on my predictions. Using these methods ultimately reduces the number of cycles of predictions, experiments, and model refinements one has to go through to achieve satisfying results.

7.4 Systems Biology for drug research

7.4.1 Comparing different approaches

7.4.1.1 Predictions based on different information

The workflow I present in this thesis proposes the application of Systems Biology methods in the field of drug development through the use of ODE models for target prediction. ODE models, however, are just one of many ways how structured biological knowledge can be used in pharmaceutical research.

Information on protein sequences and structures can be used in the search for good drug targets. In the development of anti-parasitic agents, proteins of the pathogen can be selected for being different to human counterparts, which makes it more likely that they can be selectively inhibited by drugs, or they can be selected to be structurally similar to proteins against which drugs have already been developed [Crowther et al., 2010].

The increased availability of genomic data has not only lead to the direct exploitation of this knowledge but also to the construction of stoichiometric models of metabolic pathways. Such knowledge is used to predict potential drug target through methods like choke point analysis [Yeh et al., 2004]. This analysis proposes those enzymes as targets which uniquely produce or consume a certain metabolite. Upon inhibitions of these targets complete pathways are disrupted as the enzymatic action cannot be compensated for. A similar idea can also be used to exploit protein-protein-interaction data to specifically disrupt information transmission in PPI networks [Csermely et al., 2005].

Apart from approaches relying on more simple input knowledge compared to ODE models, also more complex information can be used in the determination of effective and safe drug targets. This includes for example stochastic models, which are preferable to ODEs when small molecule numbers and stochastic fluctuations play a role in the model's behaviour, or spatial models, that might be required whenever the location of a particular substance within a compartment becomes a noteworthy factor.

7.4.1.2 Advantages of ODE models

Compared to the aforementioned approaches, ODE models involve a medium degree of complexity. On the one hand, they require experimental measurements of the dynamic behaviour of a process in order transform this knowledge into new, quantitative predictions. This makes ODE models harder to obtain than network models or protein sequence data. On the other hand,

7.4. SYSTEMS BIOLOGY FOR DRUG RESEARCH

they do not require information on the cellular localisation of model elements and are therefore more easy to construct than spatial models.

For the additional knowledge which has to be put into the construction of an ODE model, ODE based target identification approaches return more comprehensive results than those based on more simple input models. First, they allow for the rating of drug targets according to the amount of inhibitor which will be needed for an effective treatment. Second, with their help one can investigate potential side effects of a treatment, which can be included into the model. Stoichiometry- and structure-based methods do not allow for this in a straightforward manner.

Although ODE models require lots of experimental data, they have become increasingly popular amongst Systems Biologists. An indicator of this fact is the strong accumulation of models in BioModels Database over the last years [Li et al., 2010]. Due to this increased availability ODE models for many pharmacologically relevant processes might already exist in public databases. In cases in which such models are not available, many tools and web resources can support researchers in their construction. Metabolic models, for example, can be built by taking the network stoichiometry from databases like KEGG [Kanehisa et al., 2008], retrieving measured parameter values from resources like BRENDA [Scheer et al., 2011], and inserting convenience kinetics [Liebermeister and Klipp, 2006] determining the model's dynamics. Through the use of such web resources, the construction of large scale models can even be fully automatised [Borger et al., 2007].

In conclusion, it can be stated that through the amount of available relevant knowledge, ODE models can be constructed without too much effort. Furthermore, they are the most simple type of models that allow for a quantitative rating of targets based on necessary inhibitor concentrations. Thus, they are in my opinion better suited for the prediction of effective and safe drug targets.

7.4.2 Limitations of predicted drugs

7.4.2.1 Potential lack of efficacy

In contrast to drugs that have been developed following the standard target-based approach, drugs predicted via the cycle of Systems Biology have already proven their efficacy in fulfilling a given objective *in vitro*. A prerequisite for a drug to work *in vivo*, however, is that the investigated objective function is a proper quantitative measure of the state of the disease. If this objective function is not properly chosen, a drug developed using the presented methods might still not be efficacious in clinical trials. Thus, af-

7.4. SYSTEMS BIOLOGY FOR DRUG RESEARCH

ter observing a lack of *in vivo* efficacy one should investigate the molecular causes of a disease again and then repeat the systematic search for drugs with a refined objective.

7.4.2.2 Potential side effects

Although it is possible to account for potential side-effects in my framework, drugs developed with my approach might still lead to unforeseen consequences in clinical trials. Depending on the molecular causes of the side-effects, these can or cannot be accounted for by my method. First, a treatment can lead to effects in the considered network, which have not been included into the objective. If such side-effects occur, they can be accounted for in the description of the healthy state and another cycle in my proposed workflow will lead to predictions avoiding the unwanted effects. Second, side-effects might occur outside of the considered model but as a direct consequence of the inhibition in it. In such cases, it should be considered whether the model and the objective function can be extended to account for the side-effects. Third, drugs designed against a certain target might bind to off-targets and lead to unforeseen results. The occurrence of this kind of side-effects is a direct problem of the target-based approach and cannot be handled by my workflow. After targets have been selected, the design of the actual drug should be performed in a way to ensure its specificity. If a drug does not achieve a high specificity, one should either choose a different target or expect side-effects of the treatment. Although the latter idea might be associated with a high risk of failure in clinical trials, promiscuous molecules can still become commercially successful drugs. An example is the cancer drug Sunitinib, which has been shown to bind more than 70 kinases [Fabian et al., 2005].

7.4.2.3 Potential ADME problems

Problems that cannot be addressed by my approach include difficulties in the **A**dministration of the drug, its **D**istribution, its **M**etabolism, and its **E**xcretion (short ADME). Compared to the problem of finding a drug that is able to bind a selective target tightly enough, the problem of increasing a drug's bioavailability is much harder [Copeland et al., 2006]. If such a problem is encountered during drug development it might render necessary to switch to alternative targets. These treatment alternative can be provided by my methods through the application of non-identifiability analysis.

7.4. SYSTEMS BIOLOGY FOR DRUG RESEARCH

7.4.2.4 Link to personal medicine

One final problem in drug development, which cannot be anticipated by the methods presented in this work, is the fact that the efficacy of a drug in clinical trials does not necessarily imply that it achieves a certain “benefit-to-risk” ratio on a population level [Eichler et al., 2011]. According to Eichler *et al.* the difference between these two outcomes has to be attributed to various factors. First, there might be a genetic diversity for certain beneficial or hazardous alleles within a population. Second, non-genetic personal factors like age or weight as well as environmental factor like stress or grapefruit juice consumption can render individuals more susceptible to certain drug effects. Finally, issues with the prescription can arise as doctors might ignore contraindications for a drug or individuals might not adhere strictly enough to dosing intervals. These problems, however, open up the possibility of including methods of Systems Biology even closer into the drug development process by integrating information on individuals. This step towards personalised medicine will in my opinion lead to safer drugs and will ultimately, through the inclusion of a vast amount of novel knowledge, simplify the development of novel treatments.

7.4. SYSTEMS BIOLOGY FOR DRUG RESEARCH

Bibliography

- F. Achcar, E.J. Kerkhoven, B.M. Bakker, M.P. Barrett, and R. Breitling. Dynamic modelling under uncertainty: The case of trypanosoma brucei energy metabolism. *PLoS Computational Biology*, 8(1):e1002352, 2012.
- C.P. Adams and V.V. Brantner. Estimating the cost of new drug development: Is it really \$802 million? *Health Affairs*, 25(2):420, 2006.
- V. Ágoston, P. Csermely, and S. Pongor. Multiple weak hits confuse complex systems: a transcriptional regulatory network as an example. *Physical Review E*, 71(5):051909, 2005.
- H. Akaike. A new look at the statistical model identification. *Automatic Control, IEEE Transactions on*, 19(6):716–723, 1974.
- A. Alanine, M. Nettekoven, E. Roberts, and A.W. Thomas. Lead generation-enhancing the success of drug discovery by investing in the hit to lead process. *Combinatorial Chemistry & High Throughput Screening*, 6(1):51–66, 2003.
- M.A. Albert, J.R. Haanstra, V. Hannaert, J. Van Roy, F.R. Opperdoes, B.M. Bakker, and P.A.M. Michels. Experimental and in Silico Analyses of Glycolytic Flux Control in Bloodstream Form Trypanosoma brucei. *Journal of Biological Chemistry*, 280(31):28306–28315, 2005.
- R. Albert, H. Jeong, and A.L. Barabási. Error and attack tolerance of complex networks. *Nature*, 406(6794):378–382, 2000. ISSN 0028-0836.
- B. Alberts, A. Johnson, J. Lewis, M. Raff, K. Roberts, and P. Walter. *Molecular biology of the cell*. Taylor & Francis, 2007.
- B.B. Aldridge, J.M. Burke, D.A. Lauffenburger, and P.K. Sorger. Physicochemical modelling of cell signalling pathways. *Nature cell biology*, 8(11):1195–1203, 2006. ISSN 1465-7392.

BIBLIOGRAPHY

- S.F. Altschul, W. Gish, W. Miller, E.W. Myers, and D.J. Lipman. Basic local alignment search tool. *Journal of molecular biology*, 215(3):403–410, 1990.
- Antithrombotic Trialists’ Collaboration. Collaborative meta-analysis of randomised trials of antiplatelet therapy for prevention of death, myocardial infarction, and stroke in high risk patients. *Bmj*, 324(7329):71–86, 2002.
- J.F. Apgar, D.K. Witmer, F.M. White, and B. Tidor. Sloppy models, parameter uncertainty, and the role of experimental design. *Molecular BioSystems*, 6(10):1890–1900, 2010.
- B. Apsel, J.A. Blair, B. Gonzalez, T.M. Nazif, M.E. Feldman, B. Aizenstein, R. Hoffman, R.L. Williams, K.M. Shokat, and Z.A. Knight. Targeted polypharmacology: discovery of dual inhibitors of tyrosine and phosphoinositide kinases. *Nature chemical biology*, 4(11):691–699, 2008. ISSN 1552-4450.
- RP Araujo, EF Petricoin, and LA Liotta. A mathematical model of combination therapy using the EGFR signaling network. *Biosystems*, 80(1):57–69, 2005. ISSN 0303-2647.
- A.M. Aronov, S. Suresh, F.S. Buckner, W.C. Van Voorhis, C.L.M.J. Verlinde, F.R. Opperdoes, W.G.J. Hol, and M.H. Gelb. Structure-based design of submicromolar, biologically active inhibitors of trypanosomatid glyceraldehyde-3-phosphate dehydrogenase. *Proceedings of the National Academy of Sciences*, 96(8):4273, 1999.
- T.T. Ashburn and K.B. Thor. Drug repositioning: identifying and developing new uses for existing drugs. *Nature Reviews Drug Discovery*, 3(8):673–683, 2004.
- M. Ashburner, C.A. Ball, J.A. Blake, D. Botstein, H. Butler, J.M. Cherry, A.P. Davis, K. Dolinski, S.S. Dwight, J.T. Eppig, et al. Gene ontology: tool for the unification of biology. *Nature genetics*, 25(1):25–29, 2000.
- M. Ashyraliyev, J. Jaeger, and J.G. Blom. Parameter estimation and determinability analysis applied to Drosophila gap gene circuits. *BMC Systems Biology*, 2(1):83, 2008. ISSN 1752-0509.
- CJ Bacchi and N. Yarlett. Effects of antagonists of polyamine metabolism on african trypanosomes. *Acta tropica*, 54(3-4):225–236, 1993.

BIBLIOGRAPHY

- G.D. Bader, M.P. Cary, and C. Sander. Pathguide: a pathway resource list. *Nucleic Acids Research*, 34(Database Issue):D504, 2006.
- A. Bairoch, L. Bougueleret, S. Altairac, V. Amendolia, A. Auchincloss, G. Argoud-Puy, K. Axelsen, D. Baratin, MC Blatter, B. Boeckmann, et al. The universal protein resource (UniProt) 2009. *Nucleic Acids Res*, 37: D169–D174, 2009.
- B.A. Bakker, P.A.M. Michels, F.R. Opperdoes, and H.V. Westerhoff. What controls glycolysis in bloodstream form trypanosoma brucei. *Journal of Biological Chemistry*, 274(21):14551–14559, 1999.
- B.M. Bakker, H.V. Westerhoff, and P.A.M. Michels. Regulation and control of compartmentalized glycolysis in bloodstream form trypanosoma brucei. *Journal of bioenergetics and biomembranes*, 27(5):513–525, 1995.
- B.M. Bakker, P.A.M. Michels, F.R. Opperdoes, and H.V. Westerhoff. Glycolysis in bloodstream form trypanosoma brucei can be understood in terms of the kinetics of the glycolytic enzymes. *Journal of Biological Chemistry*, 272(6):3207, 1997.
- B.M. Bakker, P.A.M. Michels, M.C. Walsh, F.R. Opperdoes, and H.V. Westerhoff. *Using metabolic control analysis to improve the selectivity and effectiveness of drugs against parasitic diseases*, chapter 17. Kluwer Academic Publishers, 2000a.
- B.M. Bakker, H.V. Westerhoff, F.R. Opperdoes, and P.A.M. Michels. Metabolic control analysis of glycolysis in trypanosomes as an approach to improve selectivity and effectiveness of drugs. *Molecular and biochemical parasitology*, 106(1):1–10, 2000b.
- B.M. Bakker, H.E. Assmus, F. Bruggeman, J.R. Haanstra, E. Klipp, and H. Westerhoff. Network-Based Selectivity of Antiparasitic Inhibitors. *Molecular Biology Reports*, 29(1):1–5, 2002.
- A. Balmain, J.C. Barrett, H. Moses, and M.J. Renan. How many mutations are required for tumorigenesis? Implications from human cancer data. *Molecular carcinogenesis*, 7(3):139–146, 1993. ISSN 1098-2744.
- F.G. Banting and C.H. Best. *The internal secretion of the pancreas*. 1922.
- H.A. Bardelmeijer, J.H. Beijnen, K.R. Brouwer, H. Rosing, W.J. Nooijen, J.H.M. Schellens, and O. van Tellingen. Increased oral bioavailability of paclitaxel by gf120918 in mice through selective modulation of p-glycoprotein. *Clinical cancer research*, 6(11):4416–4421, 2000.

BIBLIOGRAPHY

- MP Barrett, DW Boykin, R. Brun, and RR Tidwell. Human african trypanosomiasis: pharmacological re-engagement with a neglected disease. *British journal of pharmacology*, 152(8):1155–1171, 2007.
- T. Barrett, D.B. Troup, S.E. Wilhite, P. Ledoux, C. Evangelista, I.F. Kim, M. Tomashevsky, K.A. Marshall, K.H. Phillippy, P.M. Sherman, R.N. Muerter, M. Holko, O. Ayanbule, A. Yefanov, and A. Soboleva. Ncbi geo: archive for functional genomics data sets–10 years on. *Nucleic Acids Research*, 39(suppl 1):D1005, 2011.
- F. Bashforth and J.C. Adams. *An attempt to test the theories of capillary action: by comparing the theoretical and measured forms of drops of fluid*. University Press, 1883.
- J. Becker and D. Kuropka. Topic-based vector space model. In *Proceedings of the 6th International Conference on Business Information Systems*, pages 7–12, 2003.
- R. Bellman. On the theory of dynamic programming. *Proceedings of the National Academy of Sciences of the United States of America*, 38(8):716, 1952.
- G. Bellu, M.P. Saccomani, S. Audoly, and L. D’Angiò. Daisy: a new software tool to test global identifiability of biological and physiological systems. *Computer methods and programs in biomedicine*, 88(1):52–61, 2007.
- A. Bender. Databases: Compound bioactivities go public. *Nature Chemical Biology*, 6(5):309–309, 2010.
- MC Berenbaum. Synergy, additivism and antagonism in immunosuppression. A critical review. *Clinical and Experimental Immunology*, 28(1):1, 1977.
- M.C. Berenbaum. The expected effect of a combination of agents: the general solution. *Journal of theoretical biology*, 114(3):413–431, 1985. ISSN 0022-5193.
- S.I. Berger and R. Iyengar. Network analyses in systems pharmacology. *Bioinformatics*, 25(19):2466, 2009.
- B.E. Bernstein, D.M. Williams, J.C. Bressi, P. Kuhn, M.H. Gelb, G.M. Blackburn, and W.G.J. Hol. A bisubstrate analog induces unexpected conformational changes in phosphoglycerate kinase from trypanosoma brucei. *Journal of molecular biology*, 279(5):1137–1148, 1998.

BIBLIOGRAPHY

- M. Berriman, E. Ghedin, C. Hertz-Fowler, G. Blandin, H. Renauld, D.C. Bartholomeu, N.J. Lennard, E. Caler, N.E. Hamlin, B. Haas, et al. The genome of the african trypanosome *trypanosoma brucei*. *Science*, 309 (5733):416, 2005.
- M.W. Berry, Z. Drmač, and E.R. Jessup. Matrices, vector spaces, and information retrieval. *SIAM review*, 41(2):335–362, 1999. ISSN 0036-1445.
- H. Bisswanger. *Enzymkinetik: Theorie und Methoden*. VCH, Weinheim, 1994.
- K.H. Bleicher, H.J. Bohm, K. Muller, and A.I. Alanine. Hit and lead generation: beyond high-throughput screening. *Nature Reviews Drug Discovery*, 2(5):369–378, 2003.
- C.I. Bliss. The toxicity of poisons applied jointly. *Annals of applied biology*, 26(3):585–615, 1939. ISSN 1744-7348.
- S. Bonhoeffer, R.M. May, G.M. Shaw, and M.A. Nowak. Virus dynamics and drug therapy. *Proceedings of the National Academy of Sciences*, 94 (13):6971, 1997.
- J.V. Bonventre. Cytosolic phospholipase a_2 α reigns supreme in arthritis and bone resorption. *Trends in immunology*, 25(3):116–119, 2004.
- B. Booth and R. Zimmel. Quest for the best. *Nature Reviews Drug Discovery*, 2(10):838–841, 2003.
- S. Borger, W. Liebermeister, J. Uhlenendorf, and E. Klipp. Automatically generated model of a metabolic network. In *International Conference on Genome Informatics*, volume 18, pages 215–224, 2007.
- A.A. Borisy, P.J. Elliott, N.W. Hurst, M.S. Lee, J. Lehár, E.R. Price, G. Serbedzija, G.R. Zimmermann, M.A. Foley, B.R. Stockwell, et al. Systematic discovery of multicomponent therapeutics. *Proceedings of the National Academy of Sciences of the United States of America*, 100(13):7977, 2003.
- A. Brazma, P. Hingamp, J. Quackenbush, G. Sherlock, P. Spellman, C. Stoeckert, J. Aach, W. Ansorge, C.A. Ball, H.C. Causton, et al. Minimum information about a microarray experiment (miame)—toward standards for microarray data. *Nature genetics*, 29(4):365–371, 2001.

BIBLIOGRAPHY

- L. Breiman and J.H. Friedman. Estimating optimal transformations for multiple regression and correlation. *Journal of the American Statistical Association*, pages 580–598, 1985.
- S. Brenner. Sequences and consequences. *Philosophical Transactions of the Royal Society B: Biological Sciences*, 365(1537):207–212, 2010.
- CG Broyden. The convergence of a class of double-rank minimization algorithms 1. general considerations. *IMA Journal of Applied Mathematics*, 6(1):76–90, 1970.
- F.J. Bruggeman and H.V. Westerhoff. The nature of systems biology. *TRENDS in Microbiology*, 15(1):45–50, 2007.
- A. Budanitsky and G. Hirst. Semantic distance in WordNet: An experimental, application-oriented evaluation of five measures. In *Workshop on WordNet and Other Lexical Resources*, volume 2, 2001.
- EC Butcher. Can cell systems biology rescue drug discovery? *Nat Rev Drug Discov*, 4:71–78, 2005.
- E.C. Butcher, E.L. Berg, and E.J. Kunkel. Systems biology in drug discovery. *Nature biotechnology*, 22(10):1253–1259, 2004.
- A.J. Cáceres, P.A.M. Michels, and V. Hannaert. Genetic validation of aldolase and glyceraldehyde-3-phosphate dehydrogenase as drug targets in trypanosoma brucei. *Molecular and biochemical parasitology*, 169(1):50–54, 2010.
- D. Calzolari, S. Bruschi, L. Coquin, J. Schofield, J.D. Feala, J.C. Reed, A.D. McCulloch, and G. Paternostro. Search algorithms as a framework for the optimization of drug combinations. *PLoS Comput Biol*, 4(12):e1000249, 2008.
- N.A. Campbell. *Biology*. Redwood City. CA: Benjamin/Cummings, 1993.
- M. Campillos, M. Kuhn, A.C. Gavin, L.J. Jensen, and P. Bork. Drug target identification using side-effect similarity. *Science*, 321(5886):263, 2008.
- G.W. Carter, P.R. Young, D.H. Albert, J. Bouska, R. Dyer, R.L. Bell, J.B. Summers, and DW Brooks. 5-lipoxygenase inhibitory activity of zileuton. *Journal of Pharmacology and Experimental Therapeutics*, 256(3):929, 1991.

BIBLIOGRAPHY

- M. Cascante, L.G. Boros, B. Comin-Anduix, P. de Atauri, J.J. Centelles, and P.W.N. Lee. Metabolic control analysis in drug discovery and disease. *Nature Biotechnology*, 20(3):243–249, 2002.
- Y.I. Cha and R.N. DuBois. Nsaids and cancer prevention: targets downstream of cox-2. *Annu. Rev. Med.*, 58:239–252, 2007.
- R. Chait, A. Craney, and R. Kishony. Antibiotic interactions that select against resistance. *Nature*, 446(7136):668–671, 2007.
- C.C. Chan, S. Boyce, C. Brideau, S. Charleson, W. Cromlish, D. Ethier, J. Evans, AW Ford-Hutchinson, MJ Forrest, JY Gauthier, et al. Rofecoxib [vioxx, mk-0966; 4-(4'-methylsulfonylphenyl)-3-phenyl-2-(5h)-furanone]: a potent and orally active cyclooxygenase-2 inhibitor. pharmacological and biochemical profiles. *Journal of Pharmacology and Experimental Therapeutics*, 290(2):551–560, 1999.
- K.H. Cheung, M. Samwald, R.K. Auerbach, and M.B. Gerstein. Structured digital tables on the semantic web: toward a structured digital literature. *Molecular Systems Biology*, 6(1), 2010.
- J. Choay, M. Petitou, JC Lormeau, P. SinaK, B. Casu, and G. Gatti. Structure-activity relationship in heparin: a synthetic pentasaccharide with high affinity for antithrombin iii and eliciting high anti-factor xa activity. *Biochemical and biophysical research communications*, 116(2):492–499, 1983.
- T.C. Chou. Drug combination studies and their synergy quantification using the chou-talalay method. *Cancer research*, 70(2):440, 2010.
- T.C. Chou and P. Talalay. Quantitative analysis of dose-effect relationships: the combined effects of multiple drugs or enzyme inhibitors. *Advances in enzyme regulation*, 22:27–55, 1984. ISSN 0065-2571.
- J.D. Clark, A.R. Schievella, E.A. Nalefski, and L.L. Lin. Cytosolic phospholipase a2. *Journal of lipid mediators and cell signalling*, 12(2-3):83, 1995.
- A.B. Clarkson, R.W. Grady, S.A. Grossman, R.J. McCallum, and F.H. Brohn. Trypanosoma brucei brucei: a systematic screening for alternatives to the salicylhydroxamic acid-glycerol combination. *Molecular and Biochemical Parasitology*, 3(5):271–291, 1981.

BIBLIOGRAPHY

- C.E. Clayton. Life without transcriptional control? from fly to man and back again. *The EMBO journal*, 21(8):1881, 2002.
- S.J. Cockell, J. Weile, P. Lord, C. Wipat, D. Andriyenko, M. Pocock, D. Wilkinson, M. Young, and A. Wipat. An integrated dataset for in silico drug discovery. *Journal of integrative bioinformatics*, 7(3), 2010. ISSN 1613-4516.
- M. Cokol, H.N. Chua, M. Tasan, B. Mutlu, Z.B. Weinstein, Y. Suzuki, M.E. Nergiz, M. Costanzo, A. Baryshnikova, G. Giaever, C. Nislow, C.L. Myers, B.J. Andrews, C. Boone, and F.P. Roth. Systematic exploration of synergistic drug pairs. *Molecular systems biology*, 7(1), 2011.
- P. Comon. Independent component analysis, a new concept? *Signal processing*, 36(3):287–314, 1994.
- R.A. Copeland, D.L. Pompliano, and T.D. Meek. Drug–target residence time and its implications for lead optimization. *Nature Reviews Drug Discovery*, 5(9):730–739, 2006.
- B. Côté, L. Boulet, C. Brideau, D. Claveau, D. Ethier, R. Frenette, M. Gagnon, A. Giroux, J. Guay, S. Guiral, et al. Substituted phenanthrene imidazoles as potent, selective, and orally active mpeg-1 inhibitors. *Bioorganic & medicinal chemistry letters*, 17(24):6816–6820, 2007a.
- R. Côté, P. Jones, L. Martens, S. Kerrien, F. Reisinger, Q. Lin, R. Leinonen, R. Apweiler, and H. Hermjakob. The protein identifier cross-referencing (picr) service: reconciling protein identifiers across multiple source databases. *BMC bioinformatics*, 8(1):401, 2007b.
- R. Courant. Variational methods for the solution of problems of equilibrium and vibrations. *Lecture notes in pure and applied mathematics*, pages 1–23, 1943.
- F.M. Couto, M.J. Silva, and P.M. Coutinho. Implementation of a functional semantic similarity measure between gene-products. Technical report, Department of Informatics, University of Lisbon, 2003.
- F.M. Couto, M.J. Silva, and P.M. Coutinho. Measuring semantic similarity between gene ontology terms. *Data & Knowledge Engineering*, 61(1):137–152, 2007.
- D. Croft, G. O Kelly, G. Wu, R. Haw, M. Gillespie, L. Matthews, M. Caudy, P. Garapati, G. Gopinath, B. Jassal, et al. Reactome: a database of

BIBLIOGRAPHY

- reactions, pathways and biological processes. *Nucleic acids research*, 39 (suppl 1):D691, 2011.
- C.N. Cronin and K.F. Tipton. Purification and regulatory properties of phosphofructokinase from trypanosoma (trypanozoon) brucei brucei. *Biochemical Journal*, 227(1):113, 1985.
- G.J. Crowther, D. Shanmugam, S.J. Carmona, M.A. Doyle, C. Hertz-Fowler, M. Berriman, S. Nwaka, S.A. Ralph, D.S. Roos, W.C. Van Voorhis, and F. Agüero. Identification of attractive drug targets in neglected-disease pathogens using an in silico approach. *PLoS neglected tropical diseases*, 4 (8):e804, 2010.
- P. Csermely, V. Ágoston, and S. Pongor. The efficiency of multi-target drugs: the network approach might help drug design. *Trends in pharmacological sciences*, 26(4):178–182, 2005. ISSN 0165-6147.
- M.S. Dasika, A. Burgard, and C.D. Maranas. A computational framework for the topological analysis and targeted disruption of signal transduction networks. *Biophysical journal*, 91(1):382–398, 2006.
- A.M. Davis, D.J. Keeling, J. Steele, N.P. Tomkinson, and A.C. Tinker. Components of successful lead generation. *Current Topics in Medicinal Chemistry*, 5(4):421–439, 2005.
- M.O. Dayhoff and R.M. Schwartz. A model of evolutionary change in proteins. In *In Atlas of protein sequence and structure*. Citeseer, 1978.
- K. Degtyarenko, P. de Matos, M. Ennis, J. Hastings, M. Zbinden, A. McNaught, R. Alcantara, M. Darsow, M. Guedj, and M. Ashburner. ChEBI: a database and ontology for chemical entities of biological interest. *Nucleic acids research*, 36(Database issue):D344, 2008.
- E. Demir, M.P. Cary, S. Paley, K. Fukuda, C. Lemer, I. Vastrik, G. Wu, P. D’Eustachio, C. Schaefer, J. Luciano, et al. The biopax community standard for pathway data sharing. *Nature biotechnology*, 28(9):935–942, 2010.
- M. Deng, K. Zhang, S. Mehta, T. Chen, and F. Sun. Prediction of protein function using protein-protein interaction data. *Journal of Computational Biology*, 10(6):947–960, 2003. ISSN 1066-5277.

BIBLIOGRAPHY

- H. Denise, C. Giroud, M.P. Barrett, and T. Baltz. Affinity chromatography using trypanocidal arsenical drugs identifies a specific interaction between glycerol-3-phosphate dehydrogenase from trypanosoma brucei and cymelarsan. *European Journal of Biochemistry*, 259(1-2):339–346, 1999.
- S. Derry and Y.K. Loke. Risk of gastrointestinal haemorrhage with long term use of aspirin: meta-analysis. *Bmj*, 321(7270):1183–1187, 2000.
- P. Deuffhard. Recent progress in extrapolation methods for ordinary differential equations. *SIAM review*, pages 505–535, 1985.
- P. Deuffhard, E. Hairer, and J. Zugck. One-step and extrapolation methods for differential-algebraic systems. *Numerische Mathematik*, 51(5):501–516, 1987.
- N. Di Gaetano, Y. Xiao, E. Erba, R. Bassan, A. Rambaldi, J. Golay, and M. Introna. Synergism between fludarabine and rituximab revealed in a follicular lymphoma cell line resistant to the cytotoxic activity of either drug alone. *British journal of haematology*, 114(4):800–809, 2001.
- L.R. Dice. Measures of the amount of ecologic association between species. *Ecology*, 26(3):297–302, 1945.
- J.A. DiMasi, R.W. Hansen, and H.G. Grabowski. The price of innovation: new estimates of drug development costs. *Journal of health economics*, 22(2):151–185, 2003.
- X.Z. Ding, C.A. Kuszynski, T.H. El-Metwally, and T.E. Adrian. Lipoxigenase inhibition induced apoptosis, morphological changes, and carbonic anhydrase expression in human pancreatic cancer cells. *Biochemical and biophysical research communications*, 266(2):392–399, 1999.
- M.E. Drew, J.C. Morris, Z. Wang, L. Wells, M. Sanchez, S.M. Landfear, and P.T. Englund. The adenosine analog tubercidin inhibits glycolysis in trypanosoma brucei as revealed by an rna interference library. *Journal of Biological Chemistry*, 278(47):46596–46600, 2003.
- B. Drewinko, TL Loo, B. Brown, JA Gottlieb, EJ Freireich, et al. Combination chemotherapy in vitro with adriamycin. observations of additive, antagonistic, and synergistic effects when used in two-drug combinations on cultured human lymphoma cells. *Cancer biochemistry biophysics*, 1(4): 187, 1976.

BIBLIOGRAPHY

- Z. Du, L. Li, C.F. Chen, P.S. Yu, and J.Z. Wang. G-SESAME: web tools for GO-term-based gene similarity analysis and knowledge discovery. *Nucleic Acids Research*, 37(suppl 2):W345, 2009. ISSN 0305-1048.
- N.C. Duarte, S.A. Becker, N. Jamshidi, I. Thiele, M.L. Mo, T.D. Vo, R. Srivas, and B.Ø. Palsson. Global reconstruction of the human metabolic network based on genomic and bibliomic data. *Proceedings of the National Academy of Sciences*, 104(6):1777, 2007.
- R.E. Durand. Synergism of cisplatin and mitomycin c in sensitive and resistant cell subpopulations of a tumor model. *International journal of cancer*, 44(5):911–917, 1989.
- J.S. Edwards, R.U. Ibarra, and B.O. Palsson. In silico predictions of escherichia coli metabolic capabilities are consistent with experimental data. *Nature biotechnology*, 19(2):125–130, 2001.
- P. Ehrlich. Chemotherapeutics: scientific principles, methods and results. *Lancet*, 2:445–451, 1913.
- H.G. Eichler, E. Abadie, A. Breckenridge, B. Flamion, L.L. Gustafsson, H. Leufkens, M. Rowland, C.K. Schneider, and B. Bloechl-Daum. Bridging the efficacy–effectiveness gap: a regulator’s perspective on addressing variability of drug response. *Nature Reviews Drug Discovery*, 10(7):495–506, 2011.
- S.M. Elbashir, J. Harborth, W. Lendeckel, A. Yalcin, K. Weber, and T. Tuschl. Duplexes of 21-nucleotide rnas mediate rna interference in cultured mammalian cells. *nature*, 411(6836):494–498, 2001.
- B. Enanga, M.R. Ariyanayagam, M.L. Stewart, and M.P. Barrett. Activity of megazol, a trypanocidal nitroimidazole, is associated with dna damage. *Antimicrobial agents and chemotherapy*, 47(10):3368, 2003.
- S.R. Engel, R. Balakrishnan, G. Binkley, K.R. Christie, M.C. Costanzo, S.S. Dwight, D.G. Fisk, J.E. Hirschman, B.C. Hitz, E.L. Hong, et al. Saccharomyces genome database provides mutant phenotype data. *Nucleic acids research*, 38(suppl 1):D433, 2010.
- K. Erguler and M.P.H. Stumpf. Practical limits for reverse engineering of dynamical systems: a statistical analysis of sensitivity and parameter inferability in systems biology models. *Mol. BioSyst.*, 2011.
- L. Euler. *Institutionum calculi integralis*, volume 1. 1768.

BIBLIOGRAPHY

- DJ Evans, PJ Barnes, SM Spaethe, EL Van Alstyne, MI Mitchell, and BJ O'connor. Effect of a leukotriene b4 receptor antagonist, ly293111, on allergen induced responses in asthma. *Thorax*, 51(12):1178–1184, 1996.
- M.A. Fabian, W.H. Biggs, D.K. Treiber, C.E. Atteridge, M.D. Azimioara, M.G. Benedetti, T.A. Carter, P. Ciceri, P.T. Edeen, M. Floyd, et al. A small molecule–kinase interaction map for clinical kinase inhibitors. *Nature biotechnology*, 23(3):329–336, 2005.
- AH Fairlamb and IBR Bowman. Trypanosoma brucei: Suramin and other trypanocidal compounds' effects on sn-glycerol-3-phosphate oxidase. *Experimental parasitology*, 43(2):353–361, 1977.
- A.A. Farooqui, H.C. Yang, T.A. Rosenberger, and L.A. Horrocks. Phospholipase a2 and its role in brain tissue. *Journal of neurochemistry*, 69(3):889–901, 1997.
- M. Farr and PA Bacon. How and when should combination therapy be used? The role of an anchor drug. *Rheumatology*, 34(suppl 2):100, 1995. ISSN 1462-0324.
- J.D. Feala, J. Cortes, P.M. Duxbury, C. Piermarocchi, A.D. McCulloch, and G. Paternostro. Systems approaches and algorithms for discovery of combinatorial therapies. *Wiley Interdisciplinary Reviews: Systems Biology and Medicine*, 2(2):181–193, 2010.
- P. Fessas, N.P. Anagnou, and D. Loukopoulos. Glycerol-3-phosphate dehydrogenase activity in the red cells of patients with thalassemia. *Blood*, 55(4):564–569, 1980.
- D.A. Fidock, P.J. Rosenthal, S.L. Croft, R. Brun, and S. Nwaka. Antimalarial drug discovery: efficacy models for compound screening. *Nature Reviews Drug Discovery*, 3(6):509–520, 2004.
- V. Fionda and L. Palopoli. Biological Network Querying Techniques: Analysis and Comparison. *Journal of Computational Biology*, 18(4):595–625, 2011. ISSN 1066-5277.
- A. Fire, S.Q. Xu, M.K. Montgomery, S.A. Kostas, S.E. Driver, and C.C. Mello. Potent and specific genetic interference by double-stranded rna in caenorhabditis elegans. *Nature*, 391(6669):806–811, 1998.
- G.A. FitzGerald. Cox-2 and beyond: approaches to prostaglandin inhibition in human disease. *Nature Reviews Drug Discovery*, 2(11):879–890, 2003.

BIBLIOGRAPHY

- G.A. FitzGerald. Coxibs and cardiovascular disease. *New England Journal of Medicine*, 351(17):1709–1711, 2004.
- J.B. Fitzgerald, B. Schoeberl, U.B. Nielsen, and P.K. Sorger. Systems biology and combination therapy in the quest for clinical efficacy. *Nature chemical biology*, 2(9):458–466, 2006.
- R. Fletcher. A new approach to variable metric algorithms. *The Computer Journal*, 13(3):317, 1970.
- R. Flindt. *Amazing numbers in biology*. Springer-Verlag Berlin Heidelberg, 2006.
- A.F. Fliri, W.T. Loging, P.F. Thadeio, and R.A. Volkmann. Analysis of drug-induced effect patterns to link structure and side effects of medicines. *Nature chemical biology*, 1(7):389–397, 2005.
- IW Flynn and IBR Bowman. The action of trypanocidal arsenical drugs on trypanosoma brucei and trypanosoma rhodesiense. *Comparative Biochemistry and Physiology Part B: Comparative Biochemistry*, 48(2):261–273, 1974.
- G. Folco and R.C. Murphy. Eicosanoid transcellular biosynthesis: from cell-cell interactions to in vivo tissue responses. *Pharmacological reviews*, 58(3):375–388, 2006.
- O. Folger, L. Jerby, C. Frezza, E. Gottlieb, E. Ruppin, and T. Shlomi. Predicting selective drug targets in cancer through metabolic networks. *Molecular Systems Biology*, 7(1), 2011.
- WN Francis and H. Kucera. The brown corpus manual: A standard corpus of present-day edited american english, for use with digital computers. *Brown University, Providence, Rhode Island*, 1964.
- W. Fujibuchi, L. Kiseleva, T. Taniguchi, H. Harada, and P. Horton. Cell-montage: similar expression profile search server. *Bioinformatics*, 23(22):3103, 2007.
- A.M. Fulton, X. Ma, and N. Kundu. Targeting prostaglandin e ep receptors to inhibit metastasis. *Cancer research*, 66(20):9794, 2006.
- C.D. Funk. Prostaglandins and leukotrienes: advances in eicosanoid biology. *Science*, 294(5548):1871, 2001.

BIBLIOGRAPHY

- C.D. Furberg, B.M. Psaty, and G.A. FitzGerald. Parecoxib, valdecoxib, and cardiovascular risk. *Circulation*, 111(3):249–249, 2005.
- Z. Gao, H. Li, H. Zhang, X. Liu, L. Kang, X. Luo, W. Zhu, K. Chen, X. Wang, and H. Jiang. PDTD: a web-accessible protein database for drug target identification. *BMC bioinformatics*, 9(1):104, 2008.
- MD Garrett and P. Workman. Discovering novel chemotherapeutic drugs for the third millennium. *European Journal of Cancer*, 35(14):2010–2030, 1999.
- S. Gay, S. Soliman, and F. Fages. A graphical method for reducing and relating models in systems biology. *Bioinformatics*, 26(18):i575, 2010. ISSN 1367-4803.
- S. Gehrig and T. Efferth. Development of drug resistance in trypanosoma brucei rhodesiense and trypanosoma brucei gambiense. treatment of human african trypanosomiasis with natural products (review). *International journal of molecular medicine*, 22(4):411, 2008.
- J. Gehrman, P.E. Hammer, C.T. Maguire, H. Wakimoto, J.K. Triedman, and C.I. Berul. Phenotypic screening for heart rate variability in the mouse. *American Journal of Physiology-Heart and Circulatory Physiology*, 279(2):H733–H740, 2000.
- A. Gelman. *Bayesian data analysis*. Texts in statistical science. Chapman & Hall/CRC, 2004. ISBN 9781584883883.
- S. Gerber, H. Aßmus, B. Bakker, and E. Klipp. Drug-efficacy depends on the inhibitor type and the target position in a metabolic network-A systematic study. *Journal of Theoretical Biology*, 252(3):442–455, 2008.
- P.K. Gessner. The isobolographic method applied to drug interactions. *Drug interactions*, pages 349–362, 1974.
- M. Ghosh, D.E. Tucker, S.A. Burchett, and C.C. Leslie. Properties of the group iv phospholipase a2 family. *Progress in lipid research*, 45(6):487–510, 2006.
- D.W. Gilroy, PR Colville-Nash, D. Willis, J. Chivers, MJ Paul-Clark, and DA Willoughby. Inducible cyclooxygenase may have anti-inflammatory properties. *Nature medicine*, 5(6):698–701, 1999.

BIBLIOGRAPHY

- K.I. Goh, M.E. Cusick, D. Valle, B. Childs, M. Vidal, and A.L. Barabási. The human disease network. *Proceedings of the National Academy of Sciences*, 104(21):8685, 2007.
- D.E. Goldberg. *Genetic algorithms in search, optimization, and machine learning*. Addison-wesley, 1989.
- D. Goldfarb. A family of variable metric methods derived by variational means. *Mathematics of Computation*, 24(109):23–26, 1970.
- M.H. Goodfellow, J. Wilson, and E. Hunt. Biochemical network matching and composition. In *Proceedings of the 2010 EDBT Workshops*, pages 1–7. ACM, 2010.
- W.R. Greco, G. Bravo, and J.C. Parsons. The search for synergy: a critical review from a response surface perspective. *Pharmacological Reviews*, 47(2):331, 1995. ISSN 0031-6997.
- GB Grindey, RG Moran, and WC Werkheiser. Approaches to the rational combination of antimetabolites for cancer chemotherapy. *Drug design*, 5: 170–249, 1975.
- CM Guldberg and P. Waage. Studier over affiniteten i. *Forhandlinger I Videnskabs-Selskabet I Christiana*, pages 35–45, 1864.
- R.N. Gutenkunst, J.J. Waterfall, F.P. Casey, K.S. Brown, C.R. Myers, and J.P. Sethna. Universally sloppy parameter sensitivities in systems biology models. *PLoS Comput Biol*, 3(10):1871–1878, 2007.
- J.R. Haanstra, A. Van Tuijl, P. Kessler, W. Reijnders, P.A.M. Michels, H.V. Westerhoff, M. Parsons, and B.M. Bakker. Compartmentation prevents a lethal turbo-explosion of glycolysis in trypanosomes. *Proceedings of the National Academy of Sciences*, 105(46):17718, 2008.
- E. Hairer and G. Wanner. *Solving ordinary differential equations II. Stiff and differential-algebraic problems*. Springer, Berlin, 1991.
- T. Handorf, M. Schulz, F. Krause, W. Jing, B. Ripkens, M. Bock, M. Flöttmann, S. Stoma, and E. Klipp. An advanced process diagram layout for biological networks. submitted, 2012.
- V. Hannaert. Sleeping sickness pathogen (*trypanosoma brucei*) and natural products: Therapeutic targets and screening systems. *Planta medica*, 77(6):586–597, 2011.

BIBLIOGRAPHY

- H. Harizi, J.B. Corcuff, and N. Gualde. Arachidonic-acid-derived eicosanoids: roles in biology and immunopathology. *Trends in molecular medicine*, 14(10):461–469, 2008.
- S. Hasan, S. Daugelat, P.S.S. Rao, and M. Schreiber. Prioritizing genomic drug targets in pathogens: application to mycobacterium tuberculosis. *PLoS Computational Biology*, 2(6):e61, 2006.
- M. Hattori, Y. Okuno, S. Goto, and M. Kanehisa. Development of a chemical structure comparison method for integrated analysis of chemical and genomic information in the metabolic pathways. *Journal of the American Chemical Society*, 125(39):11853–11865, 2003.
- M. Hausser. The hodgkin-huxley theory of the action potential. *Nature neuroscience*, 3:1165–1165, 2000.
- M. Hegreness, N. Shoresh, D. Damian, D. Hartl, and R. Kishony. Accelerated evolution of resistance in multidrug environments. *Proceedings of the National Academy of Sciences*, 105(37):13977, 2008.
- R. Heinrich and TA Rapoport. A linear steady-state treatment of enzymatic chains. General properties, control and effector strength. *Eur J Biochem*, 42(1):89–95, 1974.
- S. Helfert, A.M. Estevez, B. Bakker, P. Michels, and C. Clayton. Roles of triosephosphate isomerase and aerobic metabolism in *Trypanosoma brucei*. *Biochem J*, 357(Pt 1):117–125, 2001.
- M. Hendlich, A. Bergner, J. Günther, and G. Klebe. Relibase: Design and Development of a Database for Comprehensive Analysis of Protein-Ligand Interactions+. *Journal of molecular biology*, 326(2):607–620, 2003.
- S. Hengl, C. Kreutz, J. Timmer, and T. Maiwald. Data-based identifiability analysis of non-linear dynamical models. *Bioinformatics*, 23(19):2612, 2007.
- Ron Henkel, Lukas Endler, Nicolas Le Novère, Andre Peters, and Dagmast Waltemath. Ranked retrieval of computational biology models. *BMC Bioinformatics*, 11(423), 2010.
- M.J. Herrgård, N. Swainston, P. Dobson, W.B. Dunn, K.Y. Arga, M. Arvas, N. Büthgen, S. Borger, R. Costenoble, M. Heinemann, M. Hucka, N. Le Novère, P. Li, W. Liebermeister, M.L. Mo, A.P. Oliveira, D. Petranovic, S. Pettifer, E. Simeonidis, K. Smallbone, I. Spasić, D. Weichart,

BIBLIOGRAPHY

- R. Brent, D.S. Broomhead, H.V. Westerhoff, B.I. Kirdar, M. Penttilä, E. Klipp, B.Ø. Palsson, U. Sauer, S.G. Oliver, P. Mendes, J. Nielsen, and D.B. Kell. A consensus yeast metabolic network reconstruction obtained from a community approach to systems biology. *Nature Biotechnology*, 26(10):1155–1160, 2008.
- H.R. Herschman, S.T. Reddy, and W. Xie. Function and regulation of prostaglandin synthase-2. *Advances in Experimental Medicine and Biology*, 407:61–66, 1997.
- A.C. Hindmarsh. Lsode and lsodi, two new initial value ordinary differential equation solvers. *ACM Signum Newsletter*, 15(4):10–11, 1980.
- A.C. Hindmarsh, P.N. Brown, K.E. Grant, S.L. Lee, R. Serban, D.E. Shumaker, and C.S. Woodward. Sundials: Suite of nonlinear and differential/algebraic equation solvers. *ACM Transactions on Mathematical Software (TOMS)*, 31(3):363–396, 2005.
- A.L. Hodgkin and A.F. Huxley. A quantitative description of membrane current and its application to conduction and excitation in nerve. *The Journal of physiology*, 117(4):500, 1952.
- H.G. Holzhütter. The principle of flux minimization and its application to estimate stationary fluxes in metabolic networks. *European Journal of Biochemistry*, 271(14):2905–2922, 2004.
- W.F. Hood, J.K. Gierse, P.C. Isakson, J.R. Kiefer, R.G. Kurumbail, K. Seibert, and J.B. Monahan. Characterization of celecoxib and valdecoxib binding to cyclooxygenase. *Molecular pharmacology*, 63(4):870, 2003.
- S. Hoops, S. Sahle, R. Gauges, C. Lee, J. Pahle, N. Simus, M. Singhal, L. Xu, P. Mendes, and U. Kummer. COPASI—a COmplex PATHway SIMulator. *Bioinformatics*, 22(24):3067, 2006.
- A.L. Hopkins. Network pharmacology: the next paradigm in drug discovery. *Nature chemical biology*, 4(11):682–690, 2008. ISSN 1552-4450.
- A.L. Hopkins and C.R. Groom. The druggable genome. *Nature Reviews Drug Discovery*, 1(9):727–730, 2002.
- A.L. Hopkins, J.S. Mason, and J.P. Overington. Can we rationally design promiscuous drugs? *Current opinion in structural biology*, 16(1):127–136, 2006.

BIBLIOGRAPHY

- J. Hornberg, F. Bruggeman, B. Bakker, and H. Westerhoff. Metabolic control analysis to identify optimal drug targets. *Systems Biological Approaches in Infectious Diseases*, pages 171–189, 2007.
- J.J. Hornberg, B. Binder, F.J. Bruggeman, B. Schoeberl, R. Heinrich, and H.V. Westerhoff. Control of MAPK signalling: from complexity to what really matters. *Oncogene*, 24:5533–5542, 2005a.
- J.J. Hornberg, F.J. Bruggeman, B. Binder, C.R. Geest, A.J.M.B. de Vaate, J. Lankelma, R. Heinrich, and H.V. Westerhoff. Principles behind the multifarious control of signal transduction. *FEBS Journal*, 272(1):244–258, 2005b.
- H.R. Hotz, P. Lorenz, R. Fischer, S. Krieger, and C. Clayton. Role of 3'-untranslated regions in the regulation of hexose transporter mRNAs in *Trypanosoma brucei*. *Molecular and biochemical parasitology*, 75(1):1–14, 1995.
- CY Huang and J.E. Ferrell. Ultrasensitivity in the mitogen-activated protein kinase cascade. *Proceedings of the National Academy of Sciences*, 93(19):10078, 1996.
- Z. Huang, F. Van Harmelen, and A. Teije. Reasoning with inconsistent ontologies. In *International Joint Conference on Artificial Intelligence*, volume 19, page 454. Citeseer, 2005.
- M. Hucka, A. Finney, H.M. Sauro, H. Bolouri, J.C. Doyle, H. Kitano, A.P. Arkin, B.J. Bornstein, D. Bray, A. Cornish-Bowden, A. A. Cuellar, S. Dronov, E.D. Gilles, M. Ginkel, V. Gor, I. I. Goryanin, W. J. Hedley, T. C. Hodgman, J.-H. Hofmeyr, P. J. Hunter, N. S. Juty, J. L. Kasberger, A. Kremling, U. Kummer, N. Le Novère, L. M. Loew, D. Lucio, P. Mendes, E. Minch, E. D. Mjolsness, Y. Nakayama, M. R. Nelson, P. F. Nielsen, T. Sakurada, J. C. Schaff, B. E. Shapiro, T. S. Shimizu, H. D. Spence, J. Stelling, K. Takahashi, M. Tomita, J. Wagner, and J. Wang. The systems biology markup language (SBML): a medium for representation and exchange of biochemical network models. *Bioinformatics*, 19(4):524–531, 2003.
- Y. Hui, G. Yang, H. Galczynski, D.J. Figueroa, C.P. Austin, N.G. Copeland, D.J. Gilbert, N.A. Jenkins, and C.D. Funk. The murine cysteinyl leukotriene 2 (cyslt2) receptor. *Journal of Biological Chemistry*, 276(50):47489–47495, 2001.
- F. Hynne, S. Danø, and P.G. Sørensen. Full-scale model of glycolysis in *Saccharomyces cerevisiae*. *Biophysical Chemistry*, 94(1):121–163, 2001.

BIBLIOGRAPHY

- F. Iorio, R. Tagliaferri, and D. Bernardo. Identifying network of drug mode of action by gene expression profiling. *Journal of Computational Biology*, 16(2):241–251, 2009.
- F. Iorio, R. Bosotti, E. Scacheri, V. Belcastro, P. Mithbaokar, R. Ferriero, L. Murino, R. Tagliaferri, N. Brunetti-Pierri, A. Isacchi, et al. Discovery of drug mode of action and drug repositioning from transcriptional responses. *Proceedings of the National Academy of Sciences*, 107(33):14621–14626, 2010.
- A.I. Ivanov, A.A. Romanovsky, et al. Prostaglandin e2 as a mediator of fever: synthesis and catabolism. *Front Biosci*, 9(2):1977–1993, 2004.
- P. Jaccard. Étude comparative de la distribution florale dans une portion des Alpes et des Jura. *Bull. Soc. Vaudoise Sci. Nat*, 37:547–579, 1901.
- A.P. Jackson. Tandem gene arrays in trypanosoma brucei: comparative phylogenomic analysis of duplicate sequence variation. *BMC evolutionary biology*, 7(1):54, 2007.
- R.C. Jackson. Amphibolic drug combinations: the design of selective antimetabolite protocols based upon the kinetic properties of multienzyme systems. *Cancer research*, 53(17):3998, 1993. ISSN 0008-5472.
- J.A. Jacquez and P. Greif. Numerical parameter identifiability and estimability: Integrating identifiability, estimability, and optimal sampling design. *Mathematical Biosciences*, 77(1-2):201–227, 1985.
- E.A. Jaffe. Cell biology of endothelial cells. *Human pathology*, 18(3):234–239, 1987.
- D. Jeffreys. *Aspirin: The Extraordinary Story of a Wonder Drug*. Bloomsbury Publishing, 2010.
- C.H. Jeong, A.M. Bode, A. Pugliese, Y.Y. Cho, H.G. Kim, J.H. Shim, Y.J. Jeon, H. Li, H. Jiang, and Z. Dong. [6]-gingerol suppresses colon cancer growth by targeting leukotriene a4 hydrolase. *Cancer research*, 69(13):5584, 2009.
- Q. Ji, P. Haase, G. Qi, P. Hitzler, and S. Stadtmüller. Radon–repair and diagnosis in ontology networks. *The Semantic Web: Research and Applications*, pages 863–867, 2009.

BIBLIOGRAPHY

- J.J. Jiang and D.W. Conrath. Semantic similarity based on corpus statistics and lexical taxonomy. In *Tenth International Conference on Research on Computational Linguistics (ROCLING X)*, 1997.
- E. Jones, T. Oliphant, and P. Peterson. Scipy: Open source scientific tools for python. <http://www.scipy.org/>, 2001.
- K.S. Jones. A Statistical Interpretation of Term Specificity and its Application in Retrieval. *Journal of Documentation*, 28:11–21, 1972.
- P. Jüni, L. Nartey, S. Reichenbach, R. Sterchi, P.A. Dieppe, and M. Egger. Risk of cardiovascular events and rofecoxib: cumulative meta-analysis. *The Lancet*, 364(9450):2021–2029, 2004.
- H. Kacser and JA Burns. The control of flux. *Symp Soc Exp Biol*, 27:65–104, 1973.
- M. Kanehisa, M. Araki, S. Goto, M. Hattori, M. Hirakawa, M. Itoh, T. Katayama, S. Kawashima, S. Okuda, T. Tokimatsu, et al. KEGG for linking genomes to life and the environment. *Nucleic acids research*, 36(suppl 1):D480, 2008. ISSN 0305-1048.
- P.D. Karp, S.M. Paley, M. Krummenacker, M. Latendresse, J.M. Dale, T.J. Lee, P. Kaipa, F. Gilham, A. Spaulding, L. Popescu, et al. Pathway tools version 13.0: integrated software for pathway/genome informatics and systems biology. *Briefings in bioinformatics*, 11(1):40–79, 2010.
- L. Kaufman and PJ Rousseeuw. Finding groups in data; an introduction to cluster analysis. *Wiley Series in Probability and Mathematical Statistics. Applied Probability and Statistics Section (EUA)*., 1990.
- T. Kawamori, N. Uchiya, S. Nakatsugi, K. Watanabe, S. Ohuchida, H. Yamamoto, T. Maruyama, K. Kondo, T. Sugimura, and K. Wakabayashi. Chemopreventive effects of ono-8711, a selective prostaglandin e receptor ep1 antagonist, on breast cancer development. *Carcinogenesis*, 22(12), 2001.
- C.T. Keith, A.A. Borisy, and B.R. Stockwell. Multicomponent therapeutics for networked systems. *Nature Reviews Drug Discovery*, 4(1):71–78, 2005.
- D.B. Kell. Systems biology, metabolic modelling and metabolomics in drug discovery and development. *Drug discovery today*, 11(23-24):1085–1092, 2006.

BIBLIOGRAPHY

- B.P. Kelley, R. Sharan, R.M. Karp, T. Sittler, D.E. Root, B.R. Stockwell, and T. Ideker. Conserved pathways within bacteria and yeast as revealed by global protein network alignment. *Proceedings of the National Academy of Sciences of the United States of America*, 100(20):11394–11399, 2003.
- B.P. Kelley, B. Yuan, F. Lewitter, R. Sharan, B.R. Stockwell, and T. Ideker. PathBLAST: a tool for alignment of protein interaction networks. *Nucleic Acids Research*, 32(Web Server Issue):W83, 2004. ISSN 0305-1048.
- S.L. Kinnings, N. Liu, N. Buchmeier, P.J. Tonge, L. Xie, and P.E. Bourne. Drug discovery using chemical systems biology: repositioning the safe medicine comtan to treat multi-drug and extensively drug resistant tuberculosis. *PLoS computational biology*, 5(7):e1000423, 2009.
- S. Kirkpatrick, C.D. Gelatt, and M.P. Vecchi. Optimization by simulated annealing. *Science*, 220(4598):671, 1983.
- T. Kirsten, A. Thor, and E. Rahm. Instance-based matching of large life science ontologies. In *Proceedings of the 4th international conference on Data integration in the life sciences*, pages 172–187. Springer-Verlag, 2007.
- R.R. Klevecz, J. Bolen, G. Forrest, and D.B. Murray. A genomewide oscillation in transcription gates DNA replication and cell cycle. *Proceedings of the National Academy of Sciences of the United States of America*, 101(5):1200, 2004.
- E. Klipp, R. Heinrich, and H.G. Holzhütter. Prediction of temporal gene expression. *European journal of biochemistry*, 269(22):5406–5413, 2002.
- E. Klipp, W. Liebermeister, C. Wierling, A. Kowald, H. Lehrach, and R. Herwig. *Systems biology: a textbook*. Wiley-VCH, Weinheim, 2009.
- A. Koeberle, U. Siemoneit, U. Bühring, H. Northoff, S. Laufer, W. Albrecht, and O. Werz. Licofelone suppresses prostaglandin e2 formation by interference with the inducible microsomal prostaglandin e2 synthase-1. *Journal of Pharmacology and Experimental Therapeutics*, 326(3):975–982, 2008.
- S. Köhler, M.H. Schulz, P. Krawitz, S. Bauer, S. Dolken, C.E. Ott, C. Mundlos, D. Horn, S. Mundlos, and P.N. Robinson. Clinical diagnostics in human genetics with semantic similarity searches in ontologies. *The American Journal of Human Genetics*, 85(4):457–464, 2009. ISSN 0002-9297.
- I. Kola and J. Landis. Can the pharmaceutical industry reduce attrition rates? *Nat Rev Drug Discov*, 3(8):711–5, 2004.

BIBLIOGRAPHY

- T. Korcsmáros, M.S. Szalay, C. Bode, I.A. Kovács, and P. Csermely. How to design multi-target drugs: Target search options in cellular networks. *Expert Opin. Drug Discov.*, 2:1–10, 2007.
- J.L. Krakow and C.C. Wang. Purification and characterization of glycerol kinase from *Trypanosoma brucei*. *Molecular and biochemical parasitology*, 43(1):17–25, 1990.
- F. Krause, J. Uhlenhof, T. Lubitz, M. Schulz, E. Klipp, and W. Liebermeister. Annotation and merging of SBML models with semanticSBML. *Bioinformatics*, 26(3):421, 2010.
- F. Krause, M. Schulz, N. Swainston, and W. Liebermeister. Sustainable model building the role of standards and biological semantics. *Methods in Enzymology: Methods in Systems Biology*, 500:371, 2011.
- C. Kreutz and J. Timmer. Systems biology: experimental design. *FEBS Journal*, 276(4):923–942, 2009.
- M. Kuhn, M. Campillos, I. Letunic, L.J. Jensen, and P. Bork. A side effect resource to capture phenotypic effects of drugs. *Molecular Systems Biology*, 6(1), 2010a.
- M. Kuhn, D. Szklarczyk, A. Franceschini, M. Campillos, C. von Mering, L.J. Jensen, A. Beyer, and P. Bork. STITCH 2: an interaction network database for small molecules and proteins. *Nucleic Acids Research*, 38 (Database issue):D552, 2010b.
- C. Laibe and N. Le Novère. MIRIAM Resources: tools to generate and resolve robust cross-references in Systems Biology. *BMC Systems Biology*, 1(1):58, 2007. ISSN 1752-0509.
- N. Le Novère. Model storage, exchange and integration. *BMC Neuroscience*, 7(Suppl 1):S11, 2006. ISSN 1471-2202. doi: 10.1186/1471-2202-7-S1-S11.
- N. Le Novère, A. Finney, M. Hucka, U.S. Bhalla, F. Campagne, J. Collado-Vides, E.J. Crampin, M. Halstead, E. Klipp, P. Mendes, P. Nielsen, H. Sauro, B. Shapiro, J.L. Snoep, H.D. Spence, and B.L. Wanner. Minimum information requested in the annotation of biochemical models (MIRIAM). *Nature biotechnology*, 23(12):1509–1515, 2005.
- N. Le Novère, B. Bornstein, A. Broicher, M. Courtot, M. Donizelli, H. Dharuri, L. Li, H. Sauro, M. Schilstra, B. Shapiro, J.L. Snoep, and M. Hucka. BioModels Database: a free, centralized database of curated,

BIBLIOGRAPHY

- published, quantitative kinetic models of biochemical and cellular systems. *Nucleic Acids Research*, 34:D689–D691, 2006.
- D.S. Lee, H. Burd, J. Liu, E. Almaas, O. Wiest, A.L. Barabasi, Z.N. Oltvai, and V. Kapatral. Comparative genome-scale metabolic reconstruction and flux balance analysis of multiple staphylococcus aureus genomes identify novel antimicrobial drug targets. *Journal of bacteriology*, 191(12):4015, 2009.
- J.H. Lee, M.H. Kim, and Y.J. Lee. Information retrieval based on conceptual distance in is-a hierarchies. *Journal of Documentation*, 49(2):188–207, 1993.
- J. Lehár, G.R. Zimmermann, A.S. Krueger, R.A. Molnar, J.T. Ledell, A.M. Heilbut, G.F. Short, L.C. Giusti, G.P. Nolan, O.A. Magid, M.S. Lee, A.A. Borisy, B.R. Stockwell, and C.T. Keith. Chemical combination effects predict connectivity in biological systems. *Molecular Systems Biology*, 3(1), 2007.
- J. Lehár, A. Krueger, G. Zimmermann, and A. Borisy. High-order combination effects and biological robustness. *Molecular Systems Biology*, 4(1), 2008.
- J. Lehár, A.S. Krueger, W. Avery, A.M. Heilbut, L.M. Johansen, E.R. Price, R.J. Rickles, G.F. Short III, J.E. Staunton, X. Jin, M.S. Lee, G.R. Zimmermann, and A.A. Borisy. Synergistic drug combinations tend to improve therapeutically relevant selectivity. *Nature Biotechnology*, 27(7):659–666, 2009. ISSN 1087-0156.
- A. Levchenko, J. Bruck, and P.W. Sternberg. Scaffold proteins may biphasically affect the levels of mitogen-activated protein kinase signaling and reduce its threshold properties. *Proceedings of the National Academy of Sciences of the United States of America*, 97(11):5818, 2000.
- C. Li, M. Donizelli, N. Rodriguez, H. Dharuri, L. Endler, V. Chelliah, L. Li, E. He, A. Henry, M.I. Stefan, et al. BioModels Database: An enhanced, curated and annotated resource for published quantitative kinetic models. *BMC systems biology*, 4(1):92, 2010. ISSN 1752-0509.
- Y. Li, Z.A. Bandar, and D. McLean. An approach for measuring semantic similarity between words using multiple information sources. *IEEE Transactions on knowledge and data engineering*, 15(4):871–882, 2003.

BIBLIOGRAPHY

- Z. Li, S. Zhang, Y. Wang, X.S. Zhang, and L. Chen. Alignment of molecular networks by integer quadratic programming. *Bioinformatics*, 23(13):1631, 2007.
- W. Liebermeister. Validity and combination of biochemical models. In *Proceedings of 3rd International ESCEC Workshop on Experimental Standard Conditions on Enzyme Characterizations*, 2008.
- W. Liebermeister and E. Klipp. Bringing metabolic networks to life: convenience rate law and thermodynamic constraints. *Theoretical Biology and Medical Modelling*, 3(1):41, 2006.
- D.W. Light and R. Warburton. Demythologizing the high costs of pharmaceutical research. *BioSocieties*, 6(1):34–50, 2011.
- D. Lin. An information-theoretic definition of similarity. In *Proceedings of the 15th International Conference on Machine Learning*, volume 1, pages 296–304, 1998.
- M.A. Lindsay. Target discovery. *Nature Reviews Drug Discovery*, 2(10):831–838, 2003.
- C.A. Lipinski. Lead-and drug-like compounds: the rule-of-five revolution. *Drug Discovery Today: Technologies*, 1(4):337–341, 2004.
- T. Liu, Y. Lin, X. Wen, R.N. Jorissen, and M.K. Gilson. BindingDB: a web-accessible database of experimentally determined protein-ligand binding affinities. *Nucleic Acids Research*, 35(Database issue):D198, 2007.
- L. Ljung and T. Glad. On global identifiability for arbitrary model parametrizations. *Automatica*, 30(2):265–276, 1994.
- C.M. Lloyd, M.D.B. Halstead, and P.F. Nielsen. CellML: its future, present and past. *Progress in Biophysics and Molecular Biology*, 85(2-3):433–450, 2004.
- C.M. Lloyd, J.R. Lawson, P.J. Hunter, and P.F. Nielsen. The CellML model repository. *Bioinformatics*, 24(18):2122, 2008.
- S. Loewe and H. Muischnek. Über Kombinationswirkungen. *Naunyn-Schmiedeberg's Archives of Pharmacology*, 114(5):313–326, 1926. ISSN 0028-1298.

BIBLIOGRAPHY

- P.W. Lord, R.D. Stevens, A. Brass, and C.A. Goble. Investigating semantic similarity measures across the Gene Ontology: the relationship between sequence and annotation. *Bioinformatics*, 19(10):1275, 2003.
- T. Lubitz, M. Schulz, E. Klipp, and W. Liebermeister. Parameter balancing in kinetic models of cell metabolism. *The Journal of Physical Chemistry B*, 2010.
- K.A. Lundeen, B. Sun, L. Karlsson, and A.M. Fourie. Leukotriene b4 receptors blt1 and blt2: expression and function in human and murine mast cells. *The Journal of Immunology*, 177(5):3439–3447, 2006.
- F. Luo, Y. Yang, C.F. Chen, R. Chang, J. Zhou, and R.H. Scheuermann. Modular organization of protein interaction networks. *Bioinformatics*, 23(2):207, 2007.
- P. Ma and R. Zimmel. From the analyst’s couch: Value of novelty? *Nature Reviews Drug Discovery*, 1(8):571–572, 2002.
- A. Ma’ayan, S.L. Jenkins, J. Goldfarb, and R. Iyengar. Network analysis of fda approved drugs and their targets. *Mount Sinai Journal of Medicine: A Journal of Translational and Personalized Medicine*, 74(1):27–32, 2007.
- R. Machné, A. Finney, S. Müller, J. Lu, S. Widder, and C. Flamm. The sbml ode solver library: a native api for symbolic and fast numerical analysis of reaction networks. *Bioinformatics*, 22(11):1406, 2006.
- Rainer Machné. personal communication, 2012.
- A. Mantovani, P. Allavena, A. Sica, and F. Balkwill. Cancer-related inflammation. *Nature*, 454(7203):436–444, 2008.
- A.J. Marcus. The role of lipids in platelet function: with particular reference to the arachidonic acid pathway. *Journal of lipid research*, 19(7):793–826, 1978.
- N.I. Markevich, J.B. Hoek, and B.N. Kholodenko. Signaling switches and bistability arising from multisite phosphorylation in protein kinase cascades. *The Journal of Cell Biology*, 164(3):353, 2004.
- P. Matar, F. Rojo, R. Cassia, G. Moreno-Bueno, S. Di Cosimo, J. Tabernero, M. Guzmán, S. Rodriguez, J. Arribas, J. Palacios, et al. Combined epidermal growth factor receptor targeting with the tyrosine kinase inhibitor gefitinib (ZD1839) and the monoclonal antibody cetuximab (IMC-C225). *Clinical cancer research*, 10(19):6487, 2004. ISSN 1078-0432.

BIBLIOGRAPHY

- M.P. Mathieu. *Parexel's Bio/Pharmaceutical Ramp; D Statistical Sourcebook 2007/2008*. Barnett Educational Services/Chi, 2007.
- L.R. Matthews, P. Vaglio, J. Reboul, H. Ge, B.P. Davis, J. Garrels, S. Vincent, and M. Vidal. Identification of potential interaction networks using sequence-based searches for conserved protein-protein interactions or "interologs". *Genome research*, 11(12):2120, 2001. ISSN 1088-9051.
- M.F. McCarty. Targeting multiple signaling pathways as a strategy for managing prostate cancer: multifocal signal modulation therapy. *Integrative cancer therapies*, 3(4):349–380, 2004.
- C. Meilicke, H. Stuckenschmidt, and A. Tamilin. Repairing ontology mappings. In *Proceedings of the 22nd national conference on Artificial intelligence*, volume 2, page 1408. Menlo Park, CA; Cambridge, MA; London; AAAI Press; MIT Press; 1999, 2007.
- L. Menten and MI Michaelis. Die kinetik der invertinwirkung. *Biochem Z*, 49:333–369, 1913.
- J.B. Michel, P.J. Yeh, R. Chait, R.C. Moellering, and R. Kishony. Drug interactions modulate the potential for evolution of resistance. *Proceedings of the National Academy of Sciences*, 105(39):14918, 2008.
- G.A. Miller and W.G. Charles. Contextual correlates of semantic similarity. *Language and cognitive processes*, 6(1):1–28, 1991.
- G.A. Miller, R. Beckwith, C. Fellbaum, D. Gross, and K.J. Miller. Introduction to wordnet: An on-line lexical database*. *International Journal of lexicography*, 3(4):235, 1990.
- N. Minagawa, Y. Yabu, K. Kita, K. Nagai, N. Ohta, K. Meguro, S. Sakajo, and A. Yoshimoto. An antibiotic, ascofuranone, specifically inhibits respiration and in vitro growth of long slender bloodstream forms of trypanosoma brucei brucei. *Molecular and biochemical parasitology*, 84(2):271–280, 1997.
- H.F. Miranda, M.M. Puig, J.C. Prieto, and G. Pinardi. Synergism between paracetamol and nonsteroidal anti-inflammatory drugs in experimental acute pain. *Pain*, 121(1):22–28, 2006.
- H.P. Morgan, I.W. McNae, M.W. Nowicki, W. Zhong, P.A.M. Michels, D.S. Auld, L.A. Fothergill-Gilmore, and M.D. Walkinshaw. The trypanocidal drug suramin and other trypan blue mimetics are inhibitors of pyruvate

BIBLIOGRAPHY

- kinases and bind to the adenosine site. *Journal of Biological Chemistry*, 286(36):31232–31240, 2011.
- L.J. Morrison, L. Marcello, and R. McCulloch. Antigenic variation in the african trypanosome: molecular mechanisms and phenotypic complexity. *Cellular microbiology*, 11(12):1724–1734, 2009.
- K. Moutselos, I. Kanaris, A. Chatziioannou, I. Maglogiannis, and F.N. Kolisis. KEGGconverter: a tool for the in-silico modelling of metabolic networks of the KEGG Pathways database. *BMC bioinformatics*, 10(1):324, 2009. ISSN 1471-2105.
- E. Murabito, K. Smallbone, J. Swinton, H.V. Westerhoff, and R. Steuer. A probabilistic approach to identify putative drug targets in biochemical networks. *Journal of The Royal Society Interface*, 8(59):880–895, 2011.
- E. Nabieva, K. Jim, A. Agarwal, B. Chazelle, and M. Singh. Whole-proteome prediction of protein function via graph-theoretic analysis of interaction maps. *Bioinformatics*, 21(suppl 1):i302, 2005. ISSN 1367-4803.
- J.C. Nacher and J.M. Schwartz. A global view of drug-therapy interactions. *BMC pharmacology*, 8(1):5, 2008.
- S.B. Needleman and C.D. Wunsch. A general method applicable to the search for similarities in the amino acid sequence of two proteins. *Journal of molecular biology*, 48(3):443–453, 1970.
- I. Neeli, Z. Liu, N. Dronadula, Z.A. Ma, and G.N. Rao. An essential role of the jak-2/stat-3/cytosolic phospholipase a2 axis in platelet-derived growth factor bb-induced vascular smooth muscle cell motility. *Journal of Biological Chemistry*, 279(44):46122–46128, 2004.
- S. Nelander, W. Wang, B. Nilsson, Q.B. She, C. Pratilas, N. Rosen, P. Genemark, and C. Sander. Models from experiments: combinatorial drug perturbations of cancer cells. *Molecular systems biology*, 4(1), 2008.
- J.A. Nelder and R. Mead. A simplex method for function minimization. *The computer journal*, 7(4):308, 1965.
- I. Ngantchou, E. Nkwengoua, Y. Nganso, B. Nyasse, C. Denier, V. Hannaert, and B. Schneider. Antitrypanosomal activity of polycarpol from *piptostigma preussi* (annonaceae). *Fitoterapia*, 80(3):188–191, 2009.

BIBLIOGRAPHY

- H. Ngô, C. Tschudi, K. Gull, and E. Ullu. Double-stranded rna induces mrna degradation in trypanosoma brucei. *Proceedings of the National Academy of Sciences*, 95(25):14687, 1998.
- A.J. Nok. Azaanthraquinone inhibits respiration and in vitro growth of long slender bloodstream forms of trypanosoma congolense. *Cell biochemistry and function*, 20(3):205–212, 2002.
- C.D. Novina, P.A. Sharp, et al. The rnai revolution. *Nature*, 430(6996):161–164, 2004.
- H. Ogata, W. Fujibuchi, S. Goto, and M. Kanehisa. A heuristic graph comparison algorithm and its application to detect functionally related enzyme clusters. *Nucleic acids research*, 28(20):4021, 2000. ISSN 0305-1048.
- B.G. Olivier and J.L. Snoep. Web-based kinetic modelling using jws online. *Bioinformatics*, 20(13):2143–2144, 2004.
- H. Parkinson, M. Kapushesky, N. Kolesnikov, G. Rustici, M. Shojatalab, N. Abeygunawardena, H. Berube, M. Dylag, I. Emam, A. Farne, E. Holloway, M. Lukk, J. Malone, R. Mani, E. Pilicheva, T.F. Rayner, F. Rezwan, A. Sharma, E. Williams, X.Z. Bradley, T. Adamusiak, M. Brandizi, T. Burdett, R. Coulson, M. Krestyaninova, P. Kurnosov, E. Maruire, S.G. Neogi, P. Rocca-Serra, N. Sonson, S. Sklyar, M. Zhao, U. Sarkans, and A. Brazma. Arrayexpress update—from an archive of functional genomics experiments to the atlas of gene expression. *Nucleic acids research*, 37 (suppl 1):D868, 2009.
- S.M. Paul, D.S. Mytelka, C.T. Dunwiddie, C.C. Persinger, B.H. Munos, S.R. Lindborg, and A.L. Schacht. How to improve r&d productivity: the pharmaceutical industry’s grand challenge. *Nature Reviews Drug Discovery*, 9 (3):203–214, 2010.
- K. Pearson. On lines and planes of closest fit to systems of points in space. *Philosophical Magazine Series 6*, 2(11):559–572, 1901.
- J. Perie, I. Riviere-Alric, C. Blonski, T. Gefflaut, N.L. de Viguerie, M. Trinquier, M. Willson, F.R. Opperdoes, and M. Callens. Inhibition of the glycolytic enzymes in the trypanosome: an approach in the development of new leads in the therapy of parasitic diseases. *Pharmacology & therapeutics*, 60(2):347–365, 1993.

BIBLIOGRAPHY

- C. Pesquita, D. Faria, A.O. Falcão, P. Lord, and F.M. Couto. Semantic similarity in biomedical ontologies. *PLoS Comput Biol*, 5(7):e1000443, 2009.
- M. Peters-Golden and W.R. Henderson Jr. Leukotrienes. *New England Journal of Medicine*, 357(18):1841–1854, 2007.
- B.J. Pettus, A. Bielawska, P. Subramanian, D.S. Wijesinghe, M. Maceyka, C.C. Leslie, J.H. Evans, J. Freiberg, P. Roddy, Y.A. Hannun, et al. Ceramide 1-phosphate is a direct activator of cytosolic phospholipase a2. *Journal of Biological Chemistry*, 279(12):11320–11326, 2004.
- L. Petzold. Automatic selection of methods for solving stiff and nonstiff systems of ordinary differential equations. *SIAM J. Sci. Stat. Comput.*, 4(1):136–148, 1983.
- R.Y. Pinter, O. Rokhlenko, D. Tsur, and M. Ziv-Ukelson. Approximate labelled subtree homeomorphism. In *Combinatorial Pattern Matching*, pages 59–73. Springer, 2004.
- R.Y. Pinter, O. Rokhlenko, E. Yeger-Lotem, and M. Ziv-Ukelson. Alignment of metabolic pathways. *Bioinformatics*, 21(16):3401, 2005. ISSN 1367-4803.
- H. Pohjanpalo. System identifiability based on the power series expansion of the solution. *Mathematical biosciences*, 41(1-2):21–33, 1978.
- K. Popper. *Logik der Forschung*. 1934.
- M.J.D. Powell. A method for nonlinear constraints in minimization problems. *Optimization*, pages 283–298, 1969.
- M.J.D. Powell. A direct search optimization method that models the objective and constraint functions by linear interpolation. In *Advances in optimization and numerical analysis: proceedings of the Sixth Workshop on Optimization and Numerical Analysis, Oaxaca, Mexico*, volume 275, page 51. Kluwer Academic Pub, 1994.
- M.N. Prichard and C. Shipman. A three-dimensional model to analyze drug-drug interactions. *Antiviral research*, 14(4):181–205, 1990.
- K.D. Pruitt, K.S. Katz, H. Sicotte, D.R. Maglott, et al. Introducing refseq and locuslink: curated human genome resources at the ncbi. *Trends in genetics: TIG*, 16(1):44, 2000.

BIBLIOGRAPHY

- R. Rada, H. Mili, E. Bicknell, and M. Blettner. Development and application of a metric on semantic nets. *Systems, Man and Cybernetics, IEEE Transactions on*, 19(1):17–30, 1989.
- M.L. Radhakrishnan and B. Tidor. Optimal drug cocktail design: methods for targeting molecular ensembles and insights from theoretical model systems. *Journal of chemical information and modeling*, 48(5):1055–1073, 2008. ISSN 1549-9596.
- O. Rådmark, O. Werz, D. Steinhilber, and B. Samuelsson. 5-lipoxygenase: regulation of expression and enzyme activity. *Trends in biochemical sciences*, 32(7):332–341, 2007.
- K. Raman, P. Rajagopalan, and N. Chandra. Flux balance analysis of mycolic acid pathway: targets for anti-tubercular drugs. *PLoS computational biology*, 1(5):e46, 2005.
- R. Randhawa, C.A. Shaffer, and J.J. Tyson. Model aggregation: a building-block approach to creating large macromolecular regulatory networks. *Bioinformatics*, 25(24):3289, 2009. ISSN 1367-4803.
- R. Randhawa, C.A. Shaffer, and J.J. Tyson. Model composition for macromolecular regulatory networks. *Computational Biology and Bioinformatics, IEEE/ACM Transactions on*, 7(2):278–287, 2010. ISSN 1545-5963.
- A. Raue, C. Kreutz, T. Maiwald, J. Bachmann, M. Schilling, U. Klingmüller, and J. Timmer. Structural and practical identifiability analysis of partially observed dynamical models by exploiting the profile likelihood. *Bioinformatics*, 25(15):1923, 2009. ISSN 1367-4803.
- A. Raue, C. Kreutz, T. Maiwald, U. Klingmüller, and J. Timmer. Addressing parameter identifiability by model-based experimentation. *Systems Biology, IET*, 5(2):120–130, 2011.
- S.T. Reddy, D.J. Wadleigh, and H.R. Herschman. Transcriptional regulation of the cyclooxygenase-2 gene in activated mast cells. *Journal of Biological Chemistry*, 275(5):3107–3113, 2000.
- P. Resnik. Using Information Content to Evaluate Semantic Similarity in a Taxonomy. In *In Proceedings of the 14th International Joint Conference on Artificial Intelligence*, 1995.

BIBLIOGRAPHY

- D.R. Rhodes, S.A. Tomlins, S. Varambally, V. Mahavisno, T. Barrette, S. Kalyana-Sundaram, D. Ghosh, A. Pandey, and A.M. Chinnaiyan. Probabilistic model of the human protein-protein interaction network. *Nature biotechnology*, 23(8):951–960, 2005. ISSN 1087-0156.
- D.S. Robinson, Q. Hamid, S. Ying, A. Tsicopoulos, J. Barkans, A.M. Bentley, C. Corrigan, S.R. Durham, and A.B. Kay. Predominant th2-like bronchoalveolar t-lymphocyte population in atopic asthma. *New England Journal of Medicine*, 326(5):298–304, 1992.
- P.N. Robinson, S. Köhler, S. Bauer, D. Seelow, D. Horn, and S. Mundlos. The Human Phenotype Ontology: a tool for annotating and analyzing human hereditary disease. *The American Journal of Human Genetics*, 83(5):610–615, 2008. ISSN 0002-9297.
- GJ Roth and P.W. Majerus. The mechanism of the effect of aspirin on human platelets. i. acetylation of a particulate fraction protein. *Journal of Clinical Investigation*, 56(3):624, 1975.
- P.M. Rothwell, F.G.R. Fowkes, J.F.F. Belch, H. Ogawa, C.P. Warlow, and T.W. Meade. Effect of daily aspirin on long-term risk of death due to cancer: analysis of individual patient data from randomised trials. *The Lancet*, 377(9759):31–41, 2011.
- H. Rubenstein and J.B. Goodenough. Contextual correlates of synonymy. *Communications of the ACM*, 8(10):627–633, 1965.
- C. Runge. Über die numerische auflösung von differentialgleichungen. *Mathematische Annalen*, 46(2):167–178, 1895.
- A.P. Russ and S. Lampel. The druggable genome: an update. *Drug discovery today*, 10(23-24):1607–1610, 2005.
- D.A. Ruths, L. Nakhleh, M.S. Iyengar, S.A.G. Reddy, and P.T. Ram. Hypothesis generation in signaling networks. *Journal of Computational Biology*, 13(9):1546–1557, 2006.
- G. Salton. *The SMART retrieval system – experiments in automatic document processing*. Prentice-Hall, Inc. Upper Saddle River, NJ, USA, 1971.
- G. Salton and M.J. McGill. *Introduction to modern information retrieval*. McGraw-Hill, Inc., New York, 1986.
- F. Sams-Dodd. Target-based drug discovery: is something wrong? *Drug discovery today*, 10(2):139–147, 2005. ISSN 1359-6446.

BIBLIOGRAPHY

- B. Samuelsson, R. Morgenstern, and P.J. Jakobsson. Membrane prostaglandin synthase-1: a novel therapeutic target. *Pharmacological reviews*, 59(3):207–224, 2007.
- M. Scheer, A. Grote, A. Chang, I. Schomburg, C. Munaretto, M. Rother, C. Söhngen, M. Stelzer, J. Thiele, and D. Schomburg. Brenda, the enzyme information system in 2011. *Nucleic Acids Research*, 39(suppl 1):D670, 2011.
- A. Schlicker and M. Albrecht. Funsimmat: a comprehensive functional similarity database. *Nucleic acids research*, 36(suppl 1):D434, 2008.
- A. Schlicker, F.S. Domingues, J. Rahnenführer, and T. Lengauer. A new measure for functional similarity of gene products based on gene ontology. *BMC bioinformatics*, 7(1):302, 2006.
- B. Schoeberl, E.A. Pace, J.B. Fitzgerald, B.D. Harms, L. Xu, L. Nie, B. Linggi, A. Kalra, V. Paragas, R. Bukhalid, et al. Therapeutically targeting erbb3: a key node in ligand-induced activation of the erbb receptor-pi3k axis. *Science's STKE*, 2(77):ra31, 2009.
- M. Schulz and E. Klipp. *Introduction to Systems Biology*, chapter 3, pages 81–95. Wiley Online Library, 2010.
- M. Schulz, J. Uhlendorf, E. Klipp, and W. Liebermeister. SBMLmerge, a System for Combining Biochemical Network Models. *Genome Informatics*, 17(1):62–71, 2006.
- M. Schulz, B. Bakker, and E. Klipp. Tide: a software for the systematic scanning of drug targets in kinetic network models. *BMC bioinformatics*, 10:344, 2009.
- M. Schulz, F. Krause, N. Le Novère, E. Klipp, and W. Liebermeister. Retrieval, alignment, and clustering of computational models based on semantic annotations. *Molecular Systems Biology*, 7(513), 2011.
- M. Schulz, E. Klipp, and W. Liebermeister. Propagating semantic information in biochemical network models. *BMC bioinformatics*, 13:18, 2012.
- S. Schuster, D.A. Fell, and T. Dandekar. A general definition of metabolic pathways useful for systematic organization and analysis of complex metabolic networks. *Nature biotechnology*, 18(3):326–332, 2000.

BIBLIOGRAPHY

- B. Schwikowski, P. Uetz, and S. Fields. A network of protein-protein interactions in yeast. *Nature biotechnology*, 18(12):1257–1261, 2000. ISSN 1087-0156.
- D. Segre, A. DeLuna, G.M. Church, and R. Kishony. Modular epistasis in yeast metabolism. *Nature genetics*, 37(1):77–83, 2004.
- K. Sekiya and H. Okuda. Selective inhibition of platelet lipoxygenase by baicalein. *Biochemical and biophysical research communications*, 105(3):1090–1095, 1982.
- C.N. Serhan, N. Chiang, and T.E. Van Dyke. Resolving inflammation: dual anti-inflammatory and pro-resolution lipid mediators. *Nature Reviews Immunology*, 8(5):349–361, 2008.
- J.L. Sevilla, V. Segura, A. Podhorski, E. Guruceaga, J.M. Mato, L.A. Martinez-Cruz, F.J. Corrales, and A. Rubio. Correlation between gene expression and GO semantic similarity. *IEEE/ACM Transactions on Computational Biology and Bioinformatics (TCBB)*, 2(4):330–338, 2005. ISSN 1545-5963.
- D.F. Shanno et al. Conditioning of quasi-newton methods for function minimization. *Mathematics of computation*, 24(111):647–656, 1970.
- R. Sharan and T. Ideker. Modeling cellular machinery through biological network comparison. *Nature Biotechnology*, 24(4):427–433, 2006. ISSN 1087-0156.
- R. Sharan, S. Suthram, R.M. Kelley, T. Kuhn, S. McCuine, P. Uetz, T. Sittler, R.M. Karp, and T. Ideker. Conserved patterns of protein interaction in multiple species. *Proceedings of the National Academy of Sciences of the United States of America*, 102(6):1974, 2005.
- R. Sharan, I. Ulitsky, and R. Shamir. Network-based prediction of protein function. *Molecular systems biology*, 3(1), 2007.
- E.R. Sharlow, D. Close, T. Shun, S. Leimgruber, R. Reed, G. Mustata, P. Wipf, J. Johnson, M. O’Neil, M. Grögl, et al. Identification of potent chemotypes targeting *Leishmania major* using a high-throughput, low-stringency, computationally enhanced, small molecule screen. *PLoS Negl Trop Dis*, 3:e540, 2009.
- T. Shlomi, D. Segal, E. Ruppin, and R. Sharan. QPath: a method for querying pathways in a protein-protein interaction network. *BMC bioinformatics*, 7(1):199, 2006. ISSN 1471-2105.

BIBLIOGRAPHY

- R. Singh, J. Xu, and B. Berger. Pairwise global alignment of protein interaction networks by matching neighborhood topology. In *Research in computational molecular biology*, pages 16–31. Springer, 2007a.
- R. Singh, J. Xu, and B. Berger. Global alignment of multiple protein interaction networks with application to functional orthology detection. *Proceedings of the National Academy of Sciences*, 105(35):12763, 2008.
- S. Singh, B.K. Malik, and D.K. Sharma. Choke point analysis of metabolic pathways in *e. histolytica*: A computational approach for drug target identification. *Bioinformatics*, 2(2):68, 2007b.
- V.K. Singh and I. Ghosh. Kinetic modeling of tricarboxylic acid cycle and glyoxylate bypass in mycobacterium tuberculosis, and its application to assessment of drug targets. *Theoretical Biology and Medical Modelling*, 3(1):27, 2006.
- M.S. Skrzypek, M.B. Arnaud, M.C. Costanzo, D.O. Inglis, P. Shah, G. Binkley, S.R. Miyasato, and G. Sherlock. New tools at the candida genome database: biochemical pathways and full-text literature search. *Nucleic acids research*, 38(suppl 1):D428, 2010.
- A.A. Spector. Arachidonic acid cytochrome p450 epoxygenase pathway. *Journal of lipid research*, 50(Supplement):S52–S56, 2009.
- M. Spitzer, E. Griffiths, K.M. Blakely, J. Wildenhain, L. Ejim, L. Rossi, G. De Pascale, J. Curak, E. Brown, M. Tyers, et al. Cross-species discovery of syncretic drug combinations that potentiate the antifungal fluconazole. *Molecular Systems Biology*, 7(1), 2011.
- P. Sridhar, T. Kahveci, and S. Ranka. An iterative algorithm for metabolic network-based drug target identification. In *Pacific Symposium on Biocomputing 2007: Maui, Hawaii, 3-7 January 2007*, page 88. World Scientific Pub Co Inc, 2006.
- D.O. Stichtenoth, S. Thorén, H. Bian, M. Peters-Golden, P.J. Jakobsson, and L.J. Crofford. Microsomal prostaglandin e synthase is regulated by proinflammatory cytokines and glucocorticoids in primary rheumatoid synovial cells. *The Journal of Immunology*, 167(1):469–474, 2001.
- E.C. Stites, P.C. Trampont, Z. Ma, and K.S. Ravichandran. Network analysis of oncogenic ras activation in cancer. *Science*, 318(5849):463, 2007.

BIBLIOGRAPHY

- J.L. Stock, K. Shinjo, J. Burkhardt, M. Roach, K. Taniguchi, T. Ishikawa, H.S. Kim, P.J. Flannery, T.M. Coffman, J.D. McNeish, et al. The prostaglandin e₂ ep1 receptor mediates pain perception and regulates blood pressure. *Journal of Clinical Investigation*, 107(3):325–332, 2001.
- R. Storn and K. Price. Differential evolution—a simple and efficient heuristic for global optimization over continuous spaces. *Journal of global optimization*, 11(4):341–359, 1997.
- K. Suckling. Phospholipase a₂ s: Developing drug targets for atherosclerosis. *Atherosclerosis*, 212(2):357–366, 2010.
- N. Swainston and P. Mendes. libAnnotationSBML: a library for exploiting SBML annotations. *Bioinformatics*, 25(17):2292, 2009.
- D.C. Swinney. Biochemical mechanisms of new molecular entities (nmes) approved by united states fda during 2001-2004: mechanisms leading to optimal efficacy and safety. *Current Topics in Medicinal Chemistry*, 6(5):461–478, 2006.
- D.C. Swinney and J. Anthony. How were new medicines discovered? *Nature Reviews Drug Discovery*, 10(7):507–519, 2011.
- A.M. Tager and A.D. Luster. Blt1 and blt2: the leukotriene b₄ receptors. *Prostaglandins, leukotrienes and essential fatty acids*, 69(2):123–134, 2003.
- N. Terada, T. Yamakoshi, M. Hasegawa, H. Tanikawa, K.I. Maesako, K. Ishikawa, and A. Konno. The effect of ramatroban (bay u 3405), a thromboxane a₂ receptor antagonist, on nasal cavity volume and minimum cross-sectional area and nasal mucosal hemodynamics after nasal mucosal allergen challenge in patients with perennial allergic rhinitis. *Acta Oto-Laryngologica*, 118(Supplement 537):32–37, 1998.
- J. Tischler, B. Lehner, and A.G. Fraser. Evolutionary plasticity of genetic interaction networks. *Nature genetics*, 40(4):390–391, 2008. ISSN 1061-4036.
- Y. Tohsato, H. Matsuda, and A. Hashimoto. A multiple alignment algorithm for metabolic pathway analysis using enzyme hierarchy. In *Proceedings of the 8th International Conference on Intelligent Systems for Molecular Biology (ISMB'00)*, pages 376–383, 2000.
- P. Trouiller, P. Olliaro, E. Torreele, J. Orbinski, R. Laing, and N. Ford. Drug development for neglected diseases: a deficient market and a public-health policy failure. *The Lancet*, 359(9324):2188–2194, 2002.

BIBLIOGRAPHY

- A. Tveito and G.T. Lines. A note on a method for determining advantageous properties of an anti-arrhythmic drug based on a mathematical model of cardiac cells. *Mathematical Biosciences*, 217(2):167–173, 2009.
- F. Ushikubi, E. Segi, Y. Sugimoto, T. Murata, T. Matsuoka, T. Kobayashi, H. Hizaki, K. Tuboi, M. Katsuyama, A. Ichikawa, et al. Impaired febrile response in mice lacking the prostaglandin e receptor subtype ep3. *Nature*, 395(6699):281–284, 1998.
- S. Vajda, H. Rabitz, E. Walter, and Y. Lecourtier. Qualitative and quantitative identifiability analysis of nonlinear chemical kinetic models. *Chemical Engineering Communications*, 83(1):191–219, 1989.
- M.P. Van Iersel, A.R. Pico, T. Kelder, J. Gao, I. Ho, K. Hanspers, B.R. Conklin, and C.T. Evelo. The BridgeDb framework: standardized access to gene, protein and metabolite identifier mapping services. *BMC bioinformatics*, 11(1):5, 2010.
- V. Van Putten, Z. Refaat, C. Dessev, S. Blaine, M. Wick, L. Butterfield, S.Y. Han, L.E. Heasley, and R.A. Nemenoff. Induction of cytosolic phospholipase a2 by oncogenic ras is mediated through the jnk and erk pathways in rat epithelial cells. *Journal of Biological Chemistry*, 276(2):1226–1232, 2001.
- C.J. van Rijsbergen. *Information Retrieval*. Butterworths, London, 2nd edition, 1979.
- G. Van Rossum. *Python reference manual*. Centrum voor Wiskunde en Informatica, 1995.
- J. Vanlier, CA Tiemann, PAJ Hilbers, and NAW van Riel. A bayesian approach to targeted experiment design. *Bioinformatics*, 28(8):1136–1142, 2012.
- A. Vazquez. Optimal drug combinations and minimal hitting sets. *BMC systems biology*, 3(1):81, 2009.
- DJ Venzon and SH Moolgavkar. A method for computing profile-likelihood-based confidence intervals. *Applied Statistics*, pages 87–94, 1988.
- J. Vera, R. Curto, M. Cascante, and N.V. Torres. Detection of potential enzyme targets by metabolic modelling and optimization: Application to a simple enzymopathy. *Bioinformatics*, 23(17):2281, 2007.

BIBLIOGRAPHY

- C.L.M.J. Verlinde, V. Hannaert, C. Blonski, M. Willson, J.J. Périé, L.A. Fothergill-Gilmore, F.R. Opperdoes, M.H. Gelb, W.G.J. Hol, and P.A.M. Michels. Glycolysis as a target for the design of new anti-trypanosome drugs. *Drug Resistance Updates*, 4(1):50–65, 2001.
- I.M. Vincent, D. Creek, D.G. Watson, M.A. Kamleh, D.J. Woods, P.E. Wong, R.J.S. Burchmore, and M.P. Barrett. A molecular mechanism for efflornithine resistance in african trypanosomes. *PLoS Pathogens*, 6(11): e1001204, 2010.
- N. Visser. *Carbohydrate metabolism in erythrocytes and trypanosomes*. PhD thesis, University of Amsterdam, 1981.
- J. von Eichborn, M.S. Murgueitio, M. Dunkel, S. Koerner, P.E. Bourne, and R. Preissner. Promiscuous: a database for network-based drug-repositioning. *Nucleic acids research*, 39(suppl 1):D1060, 2011.
- D. Wang and R.N. DuBois. Eicosanoids and cancer. *Nature Reviews Cancer*, 10(3):181–193, 2010.
- H. Wang, F. Azuaje, O. Bodenreider, and J. Dopazo. Gene expression correlation and gene ontology-based similarity: an assessment of quantitative relationships. In *Computational Intelligence in Bioinformatics and Computational Biology, 2004. CIBCB'04. Proceedings of the 2004 IEEE Symposium on*, pages 25–31. IEEE, 2004.
- J.Z. Wang, Z. Du, R. Payattakool, P.S. Yu, and C.F. Chen. A new method to measure the semantic similarity of GO terms. *Bioinformatics*, 23(10): 1274, 2007. ISSN 1367-4803.
- Y. Wang, E. Bolton, S. Dracheva, K. Karapetyan, B.A. Shoemaker, T.O. Suzek, J. Wang, J. Xiao, J. Zhang, and S.H. Bryant. An overview of the pubchem bioassay resource. *Nucleic acids research*, 38(suppl 1):D255, 2010a.
- Y.T. Wang, Y.H. Huang, Y.C. Chen, C.L. Hsu, and U.C. Yang. PINT: Pathways INtegration Tool. *Nucleic Acids Research*, 2010b. ISSN 0305-1048.
- J.L. Webb. Enzyme and metabolic inhibitors. Volume 1. General principles of inhibition. *Enzyme and metabolic inhibitors. Volume 1. General principles of inhibition.*, 1963.

BIBLIOGRAPHY

- C.L. Weller, S.J. Collington, J.K. Brown, H.R.P. Miller, A. Al-Kashi, P. Clark, P.J. Jose, A. Hartnell, and T.J. Williams. Leukotriene b₄, an activation product of mast cells, is a chemoattractant for their progenitors. *The Journal of experimental medicine*, 201(12):1961, 2005.
- S.E. Wenzel. Arachidonic acid metabolites: mediators of inflammation in asthma. *Pharmacotherapy: The Journal of Human Pharmacology and Drug Therapy*, 17(1P2):3S–12S, 1997.
- A.M. Westley and J. Westley. Enzyme inhibition in open systems. *Journal of Biological Chemistry*, 271(10):5347, 1996.
- ED Wills and A. Wormall. Studies on suramin. 9. the action of the drug on some enzymes. *Biochemical Journal*, 47(2):158, 1950.
- D.S. Wishart, C. Knox, A.C. Guo, D. Cheng, S. Shrivastava, D. Tzur, B. Gautam, and M. Hassanali. DrugBank: a knowledgebase for drugs, drug actions and drug targets. *Nucleic Acids Research*, 36(Database issue):D901, 2008.
- D. Wong, M. Wang, Y. Cheng, and G.A. FitzGerald. Cardiovascular hazard and non-steroidal anti-inflammatory drugs. *Current opinion in pharmacology*, 5(2):204–210, 2005.
- S.K.M. Wong, W. Ziarko, V.V. Raghavan, and PCN Wong. On modeling of information retrieval concepts in vector spaces. *ACM Transactions on Database Systems (TODS)*, 12(2):299–321, 1987.
- D.F. Woodward, D.J. Pepperl, T.H. Burkey, and J.W. Regan. 6-isopropoxy-9-oxoxanthene-2-carboxylic acid (ah 6809), a human ep₂ receptor antagonist. *Biochemical pharmacology*, 50(10):1731–1733, 1995.
- DF Woodward, A.H.P. Krauss, J. Chen, RK Lai, CS Spada, RM Burk, SW Andrews, L. Shi, Y. Liang, KM Kedzie, et al. The pharmacology of bimatoprost (lumigan (tm)). *Survey of Ophthalmology*, 45:S337–S345, 2001.
- Z. Wu, X.M. Zhao, and L. Chen. A systems biology approach to identify effective cocktail drugs. *BMC Systems Biology*, 4(Suppl 2):S7, 2010.
- M.P. Wymann and R. Schneiter. Lipid signalling in disease. *Nature Reviews Molecular Cell Biology*, 9(2):162–176, 2008.

BIBLIOGRAPHY

- H. Xing and T.S. Gardner. The mode-of-action by network identification (mni) algorithm: a network biology approach for molecular target identification. *Nature Protocols*, 1(6):2551–2554, 2006.
- D. Xu, S.E. Rowland, P. Clark, A. Giroux, B. Côté, S. Guiral, M. Salem, Y. Ducharme, R.W. Friesen, N. Méthot, et al. Mf63 [2-(6-chloro-1h-phenanthro [9, 10-d] imidazol-2-yl)-isophthalonitrile], a selective microsomal prostaglandin e synthase-1 inhibitor, relieves pyresis and pain in preclinical models of inflammation. *Journal of Pharmacology and Experimental Therapeutics*, 326(3):754–763, 2008.
- K. Yang, W. Ma, H. Liang, Q. Ouyang, C. Tang, and L. Lai. Dynamic simulations on the arachidonic acid metabolic network. *PLoS Computational Biology*, 3(3):e55, 2007.
- K. Yang, H. Bai, Q. Ouyang, L. Lai, and C. Tang. Finding multiple target optimal intervention in disease-related molecular network. *Molecular Systems Biology*, 4(228), 2008.
- Q. Yang and S.H. Sze. Path matching and graph matching in biological networks. *Journal of Computational Biology*, 14(1):56–67, 2007. ISSN 1066-5277.
- C. Yao, D. Sakata, Y. Esaki, Y. Li, T. Matsuoka, K. Kuroiwa, Y. Sugimoto, and S. Narumiya. Prostaglandin e2–ep4 signaling promotes immune inflammation through th1 cell differentiation and th17 cell expansion. *Nature medicine*, 15(6):633–640, 2009.
- JC Yarrow, Y. Feng, ZE Perlman, T. Kirchhausen, and TJ Mitchison. Phenotypic screening of small molecule libraries by high throughput cell imaging. *Combinatorial Chemistry & High Throughput Screening*, 6(4):279–286, 2003.
- I. Yeh, T. Hanekamp, S. Tsoka, P.D. Karp, and R.B. Altman. Computational analysis of plasmodium falciparum metabolism: organizing genomic information to facilitate drug discovery. *Genome research*, 14(5):917, 2004.
- P. Yeh and R. Kishony. Networks from drug–drug surfaces. *Molecular Systems Biology*, 3(1), 2007.
- P. Yeh, A.I. Tschumi, and R. Kishony. Functional classification of drugs by properties of their pairwise interactions. *Nature genetics*, 38(4):489–494, 2006.

BIBLIOGRAPHY

- P.J. Yeh, M.J. Hegreness, A.P. Aiden, and R. Kishony. Drug interactions and the evolution of antibiotic resistance. *Nature Reviews Microbiology*, 7(6):460–466, 2009. ISSN 1740-1526.
- M.A. Yildirim, K.I. Goh, M.E. Cusick, A.L. Barabasi, and M. Vidal. Drug-target network. *Nature biotechnology*, 25(10):1119, 2007.
- B. Young, J.S. Lowe, A. Stevens, and J.W. Heath. *Wheater’s Functional Histology*. Churchill Livingstone London, 2006.
- H. Yu, N.M. Luscombe, H.X. Lu, X. Zhu, Y. Xia, J.D.J. Han, N. Bertin, S. Chung, M. Vidal, and M. Gerstein. Annotation transfer between genomes: protein–protein interologs and protein–DNA regulogs. *Genome research*, 14(6):1107, 2004. ISSN 1088-9051.
- H. Yu, R. Jansen, G. Stolovitzky, and M. Gerstein. Total ancestry measure: quantifying the similarity in tree-like classification, with genomic applications. *Bioinformatics*, 23(16):2163, 2007.
- Y. Yu, E. Ricciotti, R. Scalia, S.Y. Tang, G. Grant, Z. Yu, G. Landesberg, I. Crichton, W. Wu, E. Puré, C.D. Funk, and G.A. FitzGerald. Vascular cox-2 modulates blood pressure and thrombosis in mice. *Science Translational Medicine*, 4(132):132ra54, 2012.
- E. Yus, T. Maier, K. Michalodimitrakis, V. Van Noort, T. Yamada, W.H. Chen, J.A.H. Wodke, M. Güell, S. Martínez, R. Bourgeois, et al. Impact of genome reduction on bacterial metabolism and its regulation. *Science*, 326(5957):1263, 2009.
- A. Zaslaver, A.E. Mayo, R. Rosenberg, P. Bashkin, H. Sberro, M. Tsalyuk, M.G. Surette, and U. Alon. Just-in-time transcription program in metabolic pathways. *Nature Genetics*, 36(5):486–491, 2004.
- D.C. Zeldin. Epoxygenase pathways of arachidonic acid metabolism. *Journal of Biological Chemistry*, 276(39):36059, 2001.
- J. Zhang, M. Aizawa, S. Amari, Y. Iwasawa, T. Nakano, and K. Nakata. Development of KiBank, a database supporting structure-based drug design. *Computational biology and chemistry*, 28(5-6):401–407, 2004.
- L.V. Zhang, O.D. King, S.L. Wong, D.S. Goldberg, A.H.Y. Tong, G. Lesage, B. Andrews, H. Bussey, C. Boone, and F.P. Roth. Motifs, themes and thematic maps of an integrated *Saccharomyces cerevisiae* interaction network. *Journal of Biology*, 4(2):6, 2005. ISSN 1475-4924.

BIBLIOGRAPHY

G.R. Zimmermann, J. Lehár, and C.T. Keith. Multi-target therapeutics: when the whole is greater than the sum of the parts. *Drug discovery today*, 12(1-2):34–42, 2007. ISSN 1359-6446.

BIBLIOGRAPHY

Appendix A

Supplementary Figures and Tables

A.1 Parameters and data

A.1. PARAMETERS AND DATA

Supplementary Table A.1: Quantitative factors f_{rts} assigned to the relations in the libSBAnnotation ontology. The numerical values were chosen ad-hoc after a series of tests and systematic evaluations.

| relation type | f_{rts} |
|---------------------------|-----------|
| is_a | .5 |
| part_of | .1 |
| has_part | .1 |
| regulates | .01 |
| positively_regulates | .01 |
| negatively_regulates | .01 |
| is_tautomer_of | .9 |
| is_enantiomer_of | .01 |
| is_conjugate_acid_of | .9 |
| is_conjugate_base_of | .9 |
| has_role | .75 |
| has_functional_parent | 0. |
| is_substituent_group_from | .01 |
| has_parent_hydride | .9 |
| encodes | .1 |
| hasFunction | .75 |
| hasProcess | .25 |
| inOrganism | .1 |
| inProcess | .25 |
| isLocated | .1 |
| isPartOf | .25 |

A.1. PARAMETERS AND DATA

Supplementary Table A.2: Contribution $f_{\text{qsm}}(\mu^q, \nu^q)$ of the biological qualifiers to the annotation similarity Eq. (3.5). Each possible pair of biological qualifiers μ^q and ν^q is scored by a value between 0 and 1. The numerical values were chosen ad-hoc after a series of tests.

| f_{qsm} | is | isDescribedBy | isVersionOf | hasVersion | isHomologTo | isPartOf | hasPart | isEncodedBy | encodes |
|------------------|----|---------------|-------------|------------|-------------|----------|---------|-------------|---------|
| is | 1. | 0. | .5 | .5 | .8 | .2 | .2 | .2 | .2 |
| isDescribedBy | 0. | 1. | 0. | 0. | 0. | 0. | 0. | 0. | 0. |
| isVersionOf | .5 | 0. | .3 | .25 | .4 | .1 | .1 | .1 | .1 |
| hasVersion | .5 | 0. | .25 | .3 | .4 | .1 | .1 | .1 | .1 |
| isHomologTo | .8 | 0. | .4 | .4 | .7 | .64 | .64 | .64 | .64 |
| isPartOf | .2 | 0. | .1 | .1 | .64 | .05 | .04 | .04 | .04 |
| hasPart | .2 | 0. | .1 | .1 | .64 | .04 | .05 | .04 | .04 |
| isEncodedBy | .2 | 0. | .1 | .1 | .64 | .04 | .04 | .5 | .04 |
| encodes | .2 | 0. | .1 | .1 | .64 | .04 | .04 | .04 | .5 |

A.1. PARAMETERS AND DATA

Supplementary Table A.3: Large set of benchmark models. Models from BioModels Database were semi-automatically classified into joint biological groups taking into account the MIRIAM annotations of their <model> elements. Some annotations, e.g. GO:0000165 (MAPKKK cascade) or the annotations for organisms, referring to the NCBI Taxonomy, would have resulted in too big clusters and were therefore ignored.

| <model> annotation | Name | Models |
|------------------------------------------------------------------------------------|------------------------------------------------------------------------------------------------------------------------|---------------------------------------------------------------------------------------------------------|
| GO:0019228 | regulation of action potential in neuron | 124, 127, 129, 130, 131, 132, 133, 134, 135, 136, 141, 142 |
| GO:0006096 kegg.pathway:sce00010 | glycolysis | 42, 51, 61, 63, 64, 70, 71, 172, 176, 177, 206, 211, 225 |
| GO:0048863 | stem cell differentiation | 203, 204, 209, 210 |
| GO:0006915 kegg.pathway:hsa04210 | apoptosis | 102, 103, 220 |
| GO:0005248 GO:0019227 GO:0005249 | voltage-gated sodium channel activity, neuronal action potential propagation, voltage-gated potassium channel activity | 20, 118, 119 |
| GO:0048511 | rhythmic process | 79, 99 |
| GO:0009755 GO:0051924 | regulation of calcium ion transport, hormone-mediated signalling | 114, 115 |
| GO:0007259 kegg.pathway:mmu04630 | JAK-STAT cascade | 93, 94, 151 |
| GO:0019236 reactome:REACT_634 | response to pheromone MAP kinase cascade | 32, 116 9, 10, 11, 14 |
| GO:0016692 | NADH peroxidase activity | 46, 143 |
| GO:0007188 | G-protein signalling, coupled to cAMP nucleotide second messenger | 128, 165 |
| GO:0045990 kegg.pathway:mmu04660 | regulation of transcription by carbon catabolites T cell receptor signalling | 65, 67 139, 140, 147, 226, 227, 230 |
| GO:0009088 kegg.pathway:map00260 | threonine biosynthetic process | 66, 68 |
| GO:0008277 | regulation of G-protein coupled receptor protein signalling pathway | 85, 86 |
| GO:0006099 kegg.pathway:ko00020 GO:0006097 reactome:REACT_1785 | glyoxylate cycle, tricarboxylic acid cycle | 218, 219, 222 |
| GO:0019722 kegg.pathway:hsa04020 kegg.pathway:map04020 | calcium-mediated signalling | 39, 43, 44, 45, 47, 57, 58, 59, 60, 81, 100, 113, 117, 145, 166, 184 |
| GO:0031684 GO:0006935 | heterotrimeric G-protein complex cycle chemotaxis | 72, 80, 82 200, 229 |
| kegg.pathway:hsa04012 | ErbB signalling pathway | 175, 223 |
| GO:0006816 | calcium ion transport | 98, 162 |
| GO:0000278 kegg.pathway:sce04111 kegg.pathway:hsa04110 reactome:REACT_152 | mitotic cell cycle | 3, 4, 5, 6, 7, 8, 56, 69, 87, 107, 109, 110, 111, 144, 150, 168, 181, 186, 187, 193, 194, 196, 207, 208 |
| kegg.pathway:hsa04660 | T cell receptor signalling | 120, 122, 123 |
| GO:0007623 kegg.pathway:hsa04710 | circadian rhythm | 16, 21, 22, 24, 25, 34, 36, 55, 73, 74, 78, 83, 89, 95, 96, 97, 160, 170, 171, 214, 216 |
| GO:0016055 GO:0007173 | Wnt receptor signalling pathway epidermal growth factor receptor signalling pathway | 149, 201 19, 33, 48, 49, 84, 161 |
| kegg.pathway:hsa04115 | p53 signalling pathway | 154, 155, 156, 157, 158, 159, 188, 189 |
| GO:0007166 | cell surface receptor linked signal transduction | 1, 2, 125 |
| GO:0046655 | folic acid metabolic process | 18, 213 |
| GO:0002028 | regulation of sodium ion transport | 54, 126 |
| kegg.pathway:hsa04350 | TGF-beta signalling pathway | 101, 112, 163, 173 |
| GO:0040029 | regulation of gene expression, epigenetic | 12, 104 |
| from small example | MAPKKK cascade | 26, 27, 28, 29 |

A.1. PARAMETERS AND DATA

Supplementary Table A.4: Sensitivity of the similarity σ_{Mo}^{TVSM} and the model retrieval ranking with respect to relation type scores. Shown are the mean and the standard deviation of similarity and rank for the retrieved models when searching for models similar to BioModel 9. Mean and standard deviation are determined in 100 trials in which each relation type score is multiplied by a Gaussian distributed random variable with mean 1 and standard deviation (SD) 0.1 or 0.5. The value of the measure and the ranking by it seem quite stable with respect to variations in the f_{rts} parameter values.

| Model | SD = .1 | | | | SD = .5 | | | |
|-------|---------|----|------------------|-------|---------|------|------------------|-------|
| | Rank | | Similarity score | | Rank | | Similarity score | |
| | Mean | SD | Mean | SD | Mean | SD | Mean | SD |
| 9 | 1 | 0 | 1 | 0 | 1 | 0 | 1 | 0 |
| 11 | 2 | 0 | .925 | .0003 | 2 | 0 | .925 | .0022 |
| 14 | 3 | 0 | .865 | .0006 | 3 | 0 | .865 | .0037 |
| 10 | 4 | 0 | .816 | .0009 | 4 | 0 | .816 | .0058 |
| 26 | 5 | 0 | .737 | .0013 | 5 | 0 | .736 | .0087 |
| 28 | 6 | 0 | .687 | .0025 | 6.08 | .392 | .685 | .0174 |
| 30 | 7 | 0 | .687 | .0025 | 7.08 | .392 | .685 | .0174 |
| 27 | 8 | 0 | .673 | .0018 | 7.92 | .392 | .672 | .0122 |
| 31 | 9 | 0 | .673 | .0018 | 8.92 | .392 | .672 | .0122 |
| 29 | 10 | 0 | .614 | .0033 | 10 | 0 | .612 | .0228 |
| 84 | 11 | 0 | .482 | .0049 | 11 | 0 | .480 | .0291 |
| 116 | 12 | 0 | .397 | .0029 | 12 | 0 | .396 | .0177 |
| 32 | 13 | 0 | .348 | .0028 | 13 | 0 | .346 | .0182 |
| 149 | 14 | 0 | .335 | .0023 | 14.09 | .286 | .333 | .0158 |
| 205 | 15 | 0 | .299 | .0079 | 15.24 | .991 | .298 | .0455 |
| 33 | 16 | 0 | .260 | .0010 | 15.84 | .367 | .259 | .0065 |
| 16 | 17 | 0 | .244 | .0027 | 17.16 | .367 | .242 | .0171 |
| 49 | 18 | 0 | .240 | .0015 | 17.75 | .639 | .239 | .0098 |
| 21 | 19 | 0 | .230 | .0025 | 19.02 | .316 | .228 | .0161 |
| 4 | 20 | 0 | .222 | .0031 | 20.16 | .463 | .220 | .0199 |

A.1. PARAMETERS AND DATA

Appendix B

Supplementary methods

B.1 Mathematical details on similarity measures

B.1.1 Statistical significance of retrieved models

One of the most interesting applications of the developed similarity measures is the retrieval of relevant models and data sets from a database. This retrieval returns a list of results with their corresponding similarity values. *Per se* these similarities do not answer the question which results are significant or not. Usually, this problem is tackled by adding p-values to the results which quantify the probability that a certain similarity occurs only by chance. Given this probability and a significance level (e.g. 0.05), results can be checked for whether they are relevant or not.

B.1.1.1 Null model for the VSM

In order to be able to test for significance, one first has to develop a null model for “random” models. Such a null model is more simple to develop for the vector space model as one does not have to regard the model element and annotation structures. For the null model it should be assumed that individual features occur independently and that they are set to 1 in the random vectors with the same probability as they occur in the BioModels Database

$$p_i = \Pr(v_i^{\text{random}} = 1) = \frac{x_i + 1}{|\mathcal{M}| + 1},$$

where x_1 is the number of models in which the corresponding BC is referred to and $|\mathcal{M}|$ is the total number of models in the database.

B.1. MATHEMATICAL DETAILS ON SIMILARITY MEASURES

Given that two models show a higher similarity than expected from comparing one model to an ensemble of random models, one can assume that this similarity did not result from the fact that both models by chance contain the same common features. Therefore, one has to assume that there is some kind of “correlation” between the mentioned features and that they occur “on purpose” in both models. This would for example be the case when two models describe the same pathway.

B.1.1.2 Bayesian estimation of a p-value

When trying to develop a closed formula for the p-value for all introduced similarity measures one is facing the problem that the formulas are too complex to allow for a simple assessment of their significance. Therefore, one can perform a random sampling of similarities between the model with which the retrieval is initiated and an ensemble of random models constructed from the null model instead. Using this ensemble of random models one can employ a Bayesian approach to estimate the p-value. Given a similarity σ between two compared models, one computes the similarities of one model to a set of x_{rand} random models of which x_{sim} have a similarity $\geq \sigma$. Under the null hypothesis, x_{sim} is binomially distributed with parameters x_{rand} and an unknown p' . Following Bayesian reasoning, assuming a uniform prior distribution for the p-value on the interval $[0, 1]$, and incorporating evidence from the computed similarities, the posterior of the p-value is beta distributed [Gelman, 2004]:

$$\text{prob}(p) \sim p^{x_{\text{sim}}} (1 - p)^{x_{\text{rand}} - x_{\text{sim}}}.$$

The mean and the standard deviation of the p-value are then given by

$$\begin{aligned} \langle p \rangle &= \frac{x_{\text{sim}} + 1}{x_{\text{rand}} + 2} \\ \sqrt{\text{var}(p)} &= \sqrt{\frac{(x_{\text{sim}} + 1)(x_{\text{rand}} - x_{\text{sim}} + 1)}{(x_{\text{rand}} + 2)^2(x_{\text{rand}} + 1)}}. \end{aligned}$$

Depending on the number of random models x_{rand} in the test the minimal achievable p-value varies. E.g. setting $x_{\text{rand}} = 998$ the minimal achievable p-value is $10^{-3} \pm 10^{-3}$. In order to be able to compute lower p-values an analytic approach has to be taken.

B.1.1.3 Analytic derivation of a p-value

As mentioned before, I have been unable to derive an analytic formula for the p-values for similarity measures discussed above. This results from the

B.1. MATHEMATICAL DETAILS ON SIMILARITY MEASURES

fact that the measures are too complex and the fact that the different BCs occur with varying probabilities in the random models. I deal with the first problem by developing a way to compute p-values for a simplified measure and then showing how this approach could be extended to more complex similarities. To deal with the second problem efficiently, I employ a dynamic programming approach [Bellman, 1952] to compute the p-values numerically.

p-value for model overlap A much simplified version of the similarity $\sigma_{\text{Mo}}^{\text{TVSM}}$ is the model overlap $\sigma_{\text{Mo}}^{\text{O}}(M, N) = v_M^T v_N$ with the additional condition $\forall_i v_{iM} \in \{0, 1\}$. W.l.o.g., I assume that BCs are sorted such that $\forall_{i \in 1..|M|} v_{iM} = 1$ and $v_{iM} = 0$ otherwise. Under these conditions, the probabilities of certain scores are easy to compute, e.g.

$$\begin{aligned} \Pr(\sigma_{\text{Mo}}^{\text{O}}(M, N) = 0) &= \prod_{i=1}^{|M|} (1 - p_i) \\ \Pr(\sigma_{\text{Mo}}^{\text{O}}(M, N) = |M|) &= \prod_{i=1}^{|M|} p_i. \end{aligned}$$

It should be noted that the product in these formulas only runs to $|M|$. This is due to the fact that only the first features in v_N can contribute to the overlap score, which can be exploited for speeding up calculations.

In general the probability distribution for overlap scores can be obtained from a convolution of Bernoulli distributions, which describe the probability of a certain feature to be in a random model. Using some tricks this convolution can be calculated by dynamic programming.

In iterative steps the dynamic programming matrix D is filled with conditional probabilities $D_{x,y} = \Pr(\sigma_{\text{Mo}}^{\text{O}} = x \mid |M| = y)$. An entry $D_{x,y}$ describes the probability of a certain overlap, given that the first model refers to y features. The anchor of the iteration is given by

$$\begin{aligned} D_{0,0} &= \Pr(\sigma_{\text{Mo}}^{\text{O}} = 0 \mid |M| = 0) = 1 \\ \forall_{x>y} : D_{x,y} &= 0. \end{aligned}$$

From this anchor one can compute all entries of the D matrix using the iteration

$$\begin{aligned} D_{x,y} &= \Pr(\sigma_{\text{Mo}}^{\text{O}} = x \mid |M| = y) \\ &= p_y \cdot \Pr(\sigma_{\text{Mo}}^{\text{O}} = x - 1 \mid |M| = y - 1) \\ &\quad + (1 - p_y) \cdot \Pr(\sigma_{\text{Mo}}^{\text{O}} = x \mid |M| = y - 1) \\ &= p_y \cdot D_{x-1,y-1} + (1 - p_y) \cdot D_{x,y-1} \end{aligned}$$

B.1. MATHEMATICAL DETAILS ON SIMILARITY MEASURES

The final p-value is then computed by the formula $p = 1 - \sum_{i=0}^{\sigma_{\text{Mo}}^{\text{O}}(M,N)-1} D_{i,|M|}$, which, as mentioned above, simplifies the computation because only a part of the matrix $x \leq \sigma_{\text{Mo}}^{\text{O}}(M, N)$ has to be computed explicitly. Therefore, the computational effort of the calculation is relatively low: $\mathcal{O}(|M| \cdot \sigma_{\text{Mo}}^{\text{O}}(M, N))$.

Searching against a database The previously computed p-value describes the probability that a certain or higher similarity is observed by chance when comparing two models (of which one is a random model). When comparing a model against a model collection and asking how probable it is to see such a similarity score in any of the comparisons, one has to calculate an extended p-value

$$p_e = 1 - (1 - p)^{|\mathcal{M}|},$$

where $|\mathcal{M}|$ is the number of models in the database and p is the p-value for the comparison to one single random model.

Incorporating the similarity of BCs The simple overlap score does not include knowledge of the similarity of BCs as e.g. $\sigma_{\text{Mo}}^{\text{TVSM}}$. In order to investigate how this knowledge affects the computation of the p-value I look at a new similarity measure $\sigma_{\text{Mo}}^{\text{OS}}(M, N) = v_M^T S v_N$. Compared to the p-value computation for the overlap two problems arise. The first is that more than the first $|M|$ features in the second model can contribute to the similarity. Thus, the computation will get less efficient as bigger parts of the D matrix have to be computed. The second problem is that the similarity can take non-integer values, which increases the number of possible scores considered in the D matrix. One can reduce the additional computational effort by two ideas. First, the S matrix has to be kept as sparse as possible, e.g. by using a threshold below which the similarities of the BCs are set to zero. Second, the number of different numerical values in S has to be reduced, e.g. to multiples of $\frac{1}{2}$, which together with the first idea reduces the number of values the similarity $\sigma_{\text{Mo}}^{\text{OS}}$ can take. Apart from these numerical problems a similar iteration step has to be applied, which results straightforward from the considerations above.

Normalisation for vector length Another feature of the measure $\sigma_{\text{Mo}}^{\text{TVSM}}$ which has not been considered yet is the normalisation to the lengths of the vectors. To investigate the effects of this normalisation on the computation of the p-value, I consider the measure $\sigma_{\text{Mo}}^{\text{OL}}(M, N) = \frac{v_M^T v_N}{\sqrt{v_M^T v_M} \sqrt{v_N^T v_N}}$. Here, the computation of the p-values becomes computationally more demanding

B.2. MATHEMATICAL DETAILS ON DRUG TARGET IDENTIFICATION

because now two random variables contribute to the similarity: the overlap and the vector length minus the overlap

$$\begin{aligned}
 X &= \sum_{i=1}^{|M|} v_{iN} \\
 Y &= \sum_{i=|M|+1}^{|A|} v_{iN}.
 \end{aligned}$$

Both variables are independent. Therefore, their joint distribution is the product of their individual distributions, which can be calculated as previously described. The final p-value can in this case be computed by summing up the probabilities of all pairs X and Y fulfilling $\frac{X}{\sqrt{X+Y}} \geq \sqrt{|M|} \sigma_{Mo}^{OL}(M, N)$.

Using the discussed extensions, a p-value for the σ_{Mo}^{TVSM} could be computed. Nevertheless, this computation would be far too time consuming for the applications discussed in this thesis.

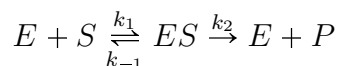
B.2 Mathematical details on drug target identification

B.2.1 Different inhibition kinetics

In the following I will introduce the kinetic formulas for various inhibition mechanisms for Michaelis-Menten (MM) kinetics. The idea behind those kinetic formulas is to simplify the reaction network, the system of differential equations, and how the system can be integrated by condensing several reaction steps acting on different time scales into a single reaction. Formulas, reaction mechanisms, and further considerations have been taken from [Bisswanger, 1994, Klipp et al., 2009].

B.2.1.1 Irreversible Michaelis-Menten kinetics

The irreversible Michaelis-Menten kinetics assumes that the binding of the substrate S to the enzyme E is a reversible process that will equilibrate rapidly, which is only valid if the substrate concentration is much higher than the enzyme concentration. Furthermore, it proposes that the enzyme-substrate complex slowly and irreversibly dissociating into enzyme and product P . An enzyme-product complex is not explicitly considered. The corresponding reaction scheme is given by



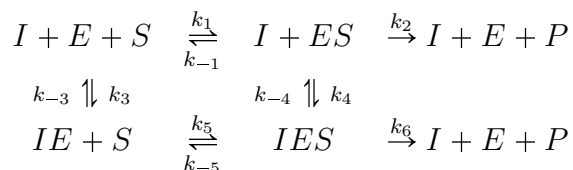
B.2. MATHEMATICAL DETAILS ON DRUG TARGET IDENTIFICATION

Using the above mentioned assumptions one can calculate the following equation for the production of P

$$\frac{dP}{dt} = V = \frac{V_{\max}S}{K_m + S},$$

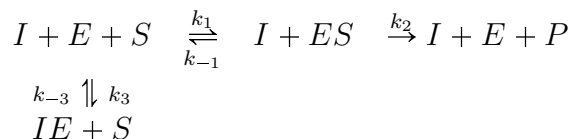
where $V_{\max} = E_{\text{tot}} \cdot k_2$ is a product of the total enzyme concentration and the enzymes catalytic rate constant and $K_m = \frac{k_{-1} + k_2}{k_1}$ ($K_m \approx \frac{k_{-1}}{k_1}$ for $k_{-1}, k_1 \gg k_2$) is the Michaelis-Menten constant. In this equation V_{\max} signifies the maximal reaction velocity for high substrate concentrations and K_m is the dissociation constant of the enzyme-substrate complex and describes the substrate concentration at which the reaction velocity reaches $\frac{V_{\max}}{2}$.

General inhibitions In any of the following inhibition types we assume that an inhibitor I can bind the enzyme (probably only in a particular state) and that it affects the reaction rate of the product formation ($k_2 \rightarrow k_6$). The general reaction scheme of an inhibitor affecting an irreversible enzymatic reaction is given by



Depending on the state of the enzyme which is bound by the inhibitor, the place where it binds it, and how it affects product formation, the reaction scheme will look different and the resulting reaction rates will vary.

Competitive inhibition The scenario in which an inhibitor binds an enzyme at the same place where the product is supposed to bind is called competitive inhibition:



An example of this scenario in nature is the inhibition of a reaction by its product. This feedback mechanism ensures that the production of P will be limited even for large concentrations of S . The resulting reaction velocity is given by

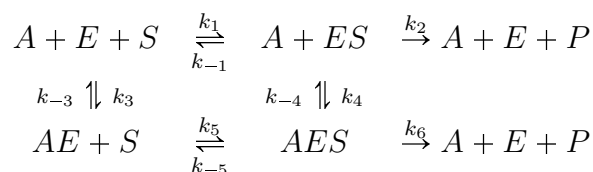
$$V = \frac{V_{\max}S}{K_m \cdot \left(1 + \frac{I}{K_I}\right) + S},$$

B.2. MATHEMATICAL DETAILS ON DRUG TARGET IDENTIFICATION

with $K_m = \frac{k_{-1}}{k_1} = \frac{k_{-5}}{k_5}$ and $K_I = \frac{k_{-3}}{k_3} = \frac{k_{-4}}{k_4}$. In this kinetic formula only the maximal reaction velocity is affected by the concentration of the inhibitor.

Other types of inhibition All of the above mentioned inhibition mechanisms assume that the binding of the inhibitor to the enzyme will prevent it from transforming the substrate to the product. In case we drop this assumption, we end up with so-called partial inhibitions. These kinetics assume that the binding of an inhibitor reduces the rate at which a product is formed but it does not completely stop the production process ($0 < k_6 < k_2$). As for the complete inhibitions different versions of partial inhibitions exist but they will not be considered in this context as they rarely occur in nature and are more of interest for theoretical considerations [Bisswanger, 1994].

Non-essential activation A very simple mechanism for the activation of an enzymatic reaction is given by the reaction scheme



Again assuming that $\frac{k_{-1}}{k_1} = \frac{k_{-5}}{k_5}$ and $\frac{k_{-3}}{k_3} = \frac{k_{-4}}{k_4}$, and furthermore supposing that $k_6 > k_2$, the reaction velocity can be written as

$$V = \frac{V_{\max} S}{K_m + S} \cdot \left(1 + \frac{A}{K_A} \right),$$

with A being the concentration of the activator, $K_m = \frac{k_{-1}}{k_1} = \frac{k_{-5}}{k_5}$, and $K_A = \frac{k_{-3}}{k_3} = \frac{k_{-4}}{k_4}$. In analogy to the noncompetitive inhibition, only the maximal reaction velocity is affected by the concentration of the activator.

B.2.1.2 Reversible Michaelis-Menten kinetics

Previously, we have only considered inhibitions to irreversible enzymatic reactions which follow Michaelis-Menten kinetics. Such reactions play a role when the enthalpy of formation of a reaction is relatively high and no substrate will be formed spontaneously from the product. For other cases the description using reversible kinetics is more suitable. The scheme of a unimolecular reversible enzymatic reaction can be drawn as



B.2. MATHEMATICAL DETAILS ON DRUG TARGET IDENTIFICATION

The velocity of a reaction following this scheme can be written as

$$\frac{dP}{dt} = V = \frac{V_{\max}^f \frac{S}{K_{mS}} - V_{\max}^r \frac{P}{K_{mP}}}{\frac{S}{K_{mS}} + \frac{P}{K_{mP}} + 1}, \quad (\text{B.1})$$

with V_{\max}^f and V_{\max}^r being the maximal reaction rates for the forward and reverse direction and K_{mS} and K_{mP} being the MM constants of the substrate and the product.

Competitive inhibition Given an inhibitor which can only bind the free enzyme and does so at the site at which substrate and product are binding, the reaction velocity changes to

$$\begin{aligned} V &= \frac{V_{\max}^f \frac{S}{K_{mS} \cdot i} - V_{\max}^r \frac{P}{K_{mP} \cdot i}}{\frac{S}{K_{mS} \cdot i} + \frac{P}{K_{mP} \cdot i} + 1} \\ &= \frac{V_{\max}^f \frac{S}{K_{mS}} - V_{\max}^r \frac{P}{K_{mP}}}{\frac{S}{K_{mS}} + \frac{P}{K_{mP}} + i}, \end{aligned}$$

with $i = 1 + \frac{I}{K_I}$. As for the competitive inhibition of the irreversible MM kinetics the inhibitor changes the apparent MM constant of the substrate and the product. In the formula this is reflected by the fact that the K_{mS}/K_{mP} values are multiplied by the factor $i = 1 + \frac{I}{K_I}$ increasing it with rising inhibitor concentrations.

Uncompetitive inhibition In cases in which the inhibitor is binding to the enzyme-substrate or the enzyme-product complex preventing their transition into each other we have an uncompetitive inhibition and the reaction velocity can be described by

$$\begin{aligned} V &= \frac{V_{\max}^f \cdot i^{-1} \frac{S}{K_{mS} \cdot i^{-1}} - V_{\max}^r \cdot i^{-1} \frac{P}{K_{mP} \cdot i^{-1}}}{\frac{S}{K_{mS} \cdot i^{-1}} + \frac{P}{K_{mP} \cdot i^{-1}} + 1} \\ &= \frac{V_{\max}^f \frac{S}{K_{mS}} - V_{\max}^r \frac{P}{K_{mP}}}{i \cdot \left(\frac{S}{K_{mS}} + \frac{P}{K_{mP}} \right) + 1}, \end{aligned}$$

with $i^{-1} = \frac{1}{1 + \frac{I}{K_I}}$. As in the irreversible case, the maximal reaction velocities and the MM constants are multiplied by the factor i^{-1} decreasing them for rising inhibitor concentrations.

B.2. MATHEMATICAL DETAILS ON DRUG TARGET IDENTIFICATION

Supplementary Table B.1: Factors to multiply the maximal reaction velocity or the MM constant with in order to to achieve different kinds of inhibitions for Michaelis-Menten type kinetics.

| inhibition type | V_{\max} | K_m |
|-----------------|-------------------------------|-------------------------------|
| competitive | 1 | $1 + \frac{I}{K_I}$ |
| uncompetitive | $\frac{1}{1 + \frac{I}{K_I}}$ | $\frac{1}{1 + \frac{I}{K_I}}$ |
| noncompetitive | $\frac{1}{1 + \frac{I}{K_I}}$ | 1 |
| activation | $1 + \frac{A}{K_A}$ | 1 |

Noncompetitive inhibition Given the scenario that the inhibitor is able to bind all forms of the enzyme, one again ends up with a noncompetitive inhibition, whose reaction velocity can be described by the formula

$$V = \frac{V_{\max}^f \frac{S}{K_{mS}} - V_{\max}^r \frac{P}{K_{mP}}}{\frac{S}{K_{mS}} + \frac{P}{K_{mP}} + 1} \cdot \frac{1}{1 + \frac{I}{K_I}}.$$

Also this kinetic formula shares a commonality with the irreversible case: only the maximal reaction velocity is affected by rising inhibitor concentrations.

Non-essential activation Similar to the behaviour of the non-essential activation, the kinetics of a non-essential activator in the reversible case are similar to those in the irreversible case. The activator increases the maximal reaction velocity by a factor $\left(1 + \frac{A}{K_A}\right)$:

$$V = \frac{V_{\max}^f \frac{S}{K_{mS}} - V_{\max}^r \frac{P}{K_{mP}}}{\frac{S}{K_{mS}} + \frac{P}{K_{mP}} + 1} \cdot \left(1 + \frac{A}{K_A}\right).$$

B.2.1.3 General considerations on inhibition/activation kinetics

As shown in this section of the Appendix, the kinetic formulas for different kinds of inhibitions do only differ by the fact which kind of variables are altered by multiplication or division with the term $1 + \frac{I}{K_I}$ (compare Table B.1). This fact is used in the TIDE tool in order to automatically create possible inhibition kinetics from a non-inhibited kinetic after the V_{\max} and the K_m variables have been identified.

B.2. MATHEMATICAL DETAILS ON DRUG TARGET IDENTIFICATION

These variables can be identified in a kinetic formula by numerically evaluating it for various changes to variable values. A V_{\max} value can be identified by the behaviour that the reaction velocity should scale linearly with it. Thus, replacing V_{\max} by $2 \cdot V_{\max}$ should change the reaction velocity v to $2 \cdot v$. For kinetics in which this fact hold for more than one variable (e.g. if V_{\max} is split into $k_{\text{cat}} \cdot E_{\text{tot}}$) we can simply select one of these variables and proceed with the aforementioned multiplication. For K_m values the identification is a little more complex as we have one value for every substrate (and for reversible reactions one for each product, too). Given that we want to identify the K_{mS_i} value for a substrate S_i , it will be the only variable x for which holds $\forall_{y \in \mathbb{R}} : v(S_i, x, \dots) = v(y \cdot S_i, y \cdot x, \dots)$. It should be noted that in case a reaction is reversible, both, the K_m of the substrate and the K_m of the product are affected by the inhibition. Furthermore, if a reversible reaction involves multiple substrates and products, an inhibition will always affect a certain substrate/product pair, which cannot necessarily be determined automatically.

B.2.2 Construction of a objective function for drug target identification

B.2.2.1 Proof: All objectives are fulfilled if the objective function has a value smaller than one

The construction of the objective function (\mathcal{X}^2) from a description of the healthy state has been discussed in 5.2.1.3 in the main text. I have proposed that $\mathcal{X}^2 < 1$ implies that the objective associated with each single concentration time point is satisfied.

Proof. Lets assume that one of the aims is not fulfilled. Depending on the type of the aim, we have three different cases.

- The minimization is not achieved if $y_i(t_j) = \bar{y}_i(t_j) + \varepsilon$ with $\varepsilon \geq 0$. In this case the corresponding summand of the χ^2 value reads

$$\begin{aligned} \left(\frac{0 - y_i(t_j)}{\bar{y}_i(t_j)} \right)^2 &= \left(\frac{-\bar{y}_i(t_j) - \varepsilon}{\bar{y}_i(t_j)} \right)^2 \\ &= \left(1 + \frac{\varepsilon}{\bar{y}_i(t_j)} \right)^2 \\ &\geq 1 \end{aligned}$$

B.2. MATHEMATICAL DETAILS ON DRUG TARGET IDENTIFICATION

- The maximization is not achieved if $y_i(t_j) = \bar{y}_i(t_j) - \varepsilon$ with $\varepsilon \geq 0$. In this case the corresponding summand of the χ^2 value reads

$$\begin{aligned} \left(\frac{0 - \frac{1}{y_i(t_j)}}{\frac{1}{\bar{y}_i(t_j)}} \right)^2 &= \left(\frac{0 - \frac{1}{\bar{y}_i(t_j) - \varepsilon}}{\frac{1}{\bar{y}_i(t_j)}} \right)^2 \\ &= \left(\frac{\bar{y}_i(t_j)}{\bar{y}_i(t_j) - \varepsilon} \right)^2 \\ &\geq 1 \end{aligned}$$

- Keeping a value in a certain range is not fulfilled, if the value lies e.g. below the lower boundary of the range $y_i(t_j) = \bar{y}_i(t_j) - \sigma_i(t_j) - \varepsilon$

$$\begin{aligned} \left(\frac{\bar{y} - y_i(t_j)}{\sigma_i(t_j)} \right)^2 &= \left(\frac{\sigma_i(t_j) + \varepsilon}{\sigma_i(t_j)} \right)^2 \\ &\geq 1 \end{aligned}$$

Considering the separate summands in the χ^2 value $\mathcal{X}^2 = \sum_i \mathcal{X}_i^2$ and assuming that one of the aims is not fulfilled, we have $\exists_i : \mathcal{X}_i^2 \geq 1$ and because of $\forall_i : \mathcal{X}_i^2 \geq 0$ it follows that $\mathcal{X}^2 \geq 1$. \square

B.2.3 Parameter identifiability

B.2.3.1 Mathematical reasoning for investigating \mathcal{X}^2

In parameter estimation problems we are given the objective to find the most likely set of parameter values given certain experimental data \bar{y} . We generally assume those data points can be described by an ODE system

$$\frac{dy(t)}{dt} = f(y, \theta, t)$$

and that the data are observations of the underlying ODE system (including the real parameter values θ^*) including measurement errors

$$\bar{y}(t) = y(t, \theta^*) + \sigma(t)\varepsilon \tag{B.2}$$

where ε is standard normally distributed random variable.

Maximum likelihood Given this assumptions the probability of the measured data set \bar{Y} given the parameters θ is expressed by

$$\Pr(\bar{Y}|\theta) = \frac{1}{\sqrt{2\pi\sigma^2}} \prod_{\bar{y} \in \bar{Y}} \exp\left(-\frac{(\bar{y}(t) - y(t, \theta))^2}{2\sigma^2}\right),$$

B.2. MATHEMATICAL DETAILS ON DRUG TARGET IDENTIFICATION

which is the likelihood of the parameters $L(\theta)$.

From this we would like to infer the behaviour of the posterior probability

$$\Pr(\theta|\bar{Y}) = \frac{\Pr(\bar{Y}|\theta) \Pr(\theta)}{\Pr(\bar{Y})}.$$

While $\Pr(\bar{Y})$ is just a normalization, $\Pr(\theta)$ is in most cases assumed to be constant over a broad parameter range, from which follows

$$\frac{\partial \Pr(\theta)}{\partial \theta} = 0$$

and thus

$$\frac{\partial \Pr(\theta|\bar{Y})}{\partial \theta} = \frac{\Pr(\theta)}{\Pr(\bar{Y})} \frac{\partial \Pr(\bar{Y}|\theta)}{\partial \theta}$$

Using this assumption we can write the sensitivity of the posterior with respect to the parameters

$$\begin{aligned} \frac{\partial \ln \Pr(\theta|\bar{Y})}{\partial \ln \theta} &= \frac{\partial \Pr(\theta|\bar{Y})}{\partial \theta} \frac{\theta}{\Pr(\theta|\bar{Y})} \\ &= \frac{\partial \Pr(\bar{Y}|\theta)}{\partial \theta} \frac{\Pr(\theta)}{\Pr(\bar{Y})} \frac{\theta \Pr(\bar{Y})}{\Pr(\bar{Y}|\theta) \Pr(\theta)} \\ &= \frac{\partial \ln \Pr(\bar{Y}|\theta)}{\partial \ln \theta}. \end{aligned} \tag{B.3}$$

Setting Eq. (B.3) to 0 one can see that the posterior has extremal values at the same parameter values as the likelihood. So, maximizing the posterior is equivalent to maximizing the likelihood.

χ^2 distribution Assuming that the normalized residuals

$$\frac{\bar{y}(t) - y(t, \theta)}{\sigma(t)}$$

are independent and standard normally distributed (compare Eq. (B.2)), we know that the sum of squared, normalized residuals is χ^2 distributed

$$\sum \left(\frac{\bar{y}(t_i) - y(t_i, \theta)}{\sigma(t_i)} \right)^2 = \chi^2.$$

B.2. MATHEMATICAL DETAILS ON DRUG TARGET IDENTIFICATION

Connection between likelihood and \mathcal{X}^2

$$\begin{aligned}
 L(\theta) &= \Pr(\bar{Y}|\theta) \\
 &= \frac{1}{\sqrt{2\pi\sigma^2}} \prod_{\bar{y} \in \bar{Y}} \exp\left(-\frac{(\bar{y}(t) - y(t, \theta))^2}{2\sigma^2}\right) \\
 &= \frac{1}{\sqrt{2\pi\sigma^2}} \exp\left(-\frac{1}{2} \sum_{\bar{y} \in \bar{Y}} \left(\frac{\bar{y}(t) - y(t, \theta)}{\sigma}\right)^2\right)
 \end{aligned}$$

from which follows

$$\begin{aligned}
 \ln L(\theta) &= \ln\left(\frac{1}{\sqrt{2\pi\sigma^2}}\right) - \frac{1}{2}\mathcal{X}^2 \\
 \Rightarrow \quad \mathcal{X}^2 &= c - 2\ln L(\theta)
 \end{aligned}$$

Since we are now aware of the relations between conditional probabilities, likelihood, and \mathcal{X}^2 , we can formulate the objective of the parameter estimation: finding a parameter set

$$\begin{aligned}
 \hat{\theta} &= \arg \max [\Pr(\theta|\bar{Y})] \\
 &= \arg \max [\ln \Pr(\theta|\bar{Y})] \\
 &= \arg \max [\ln L(\theta)] \\
 &= \arg \min [\mathcal{X}^2(\theta)]
 \end{aligned}$$

B.2.3.2 Rules for interchangeabilities

Given a working treatment \mathcal{I} with a list of corresponding drug concentrations $i_{\mathcal{I}}$ I propose the following theorems for treatments and interchangeabilities:

- $T(\mathcal{I}) \Rightarrow T(\mathcal{I} + \{x\})$

Proof.

$$\begin{aligned}
 T(\mathcal{I}) &\Rightarrow \mathcal{X}^2(\theta_{\mathcal{I}}) < 1 \\
 &\Rightarrow \exists_{\epsilon > 0} \mathcal{X}^2(\theta_{\mathcal{I}}) + \epsilon < 1
 \end{aligned}$$

Using $\theta_x = (\theta_0 = 0, \dots, \theta_x = 1, \dots, \theta_n = 0)$ and knowing that the objective function is continuous in the parameter space we know that

$$\exists_{\delta > 0} \mathcal{X}^2(\theta_{\mathcal{I}} + \delta\theta_x) \leq \mathcal{X}^2(\theta_{\mathcal{I}}) + \epsilon.$$

B.2. MATHEMATICAL DETAILS ON DRUG TARGET IDENTIFICATION

Inserting this we end up with

$$\begin{aligned} \exists_{\epsilon>0} \mathcal{X}^2(\theta_{\mathcal{I}}) + \epsilon < 1 &\Rightarrow \exists_{\delta>0} \mathcal{X}^2(\theta_{\mathcal{I}} + \delta\theta_x) < 1 \\ &\Rightarrow T(\mathcal{I} + \{x\}) \end{aligned}$$

□

- $T(\mathcal{I}) \Rightarrow \exists_{i_x>0, i_y>0} I(i_x \leftrightarrow i_y | i_{\mathcal{I}})$

Proof. From $T(\mathcal{I})$ one concludes that $T(\mathcal{I} + \{x\})$, $T(\mathcal{I} + \{y\})$, and $T(\mathcal{I} + \{x, y\})$. Along small values for i_x and i_y the \mathcal{X}^2 value does not change significantly (see proof above), and thus we can conclude interchangeability for them: $\exists_{i_x>0, i_y>0} I(i_x \leftrightarrow i_y | i_{\mathcal{I}})$. □

- $I(i_x \rightarrow i_y | i_{\mathcal{I}}) \Rightarrow \forall_z \exists_{i_z} I(i_x \rightarrow i_y, i_z | i_{\mathcal{I}})$
- $I(i_x \rightarrow i_y | i_{\mathcal{I}}) \& I(i_y \rightarrow i_z | i_{\mathcal{I}}) \Rightarrow I(i_x \rightarrow i_z | i_{\mathcal{I}})$
- $I(i_x \leftrightarrow i_y | i_{\mathcal{I}}) \& I(i_y \leftrightarrow i_z | i_{\mathcal{I}}) \Rightarrow I(i_x \leftrightarrow i_z | i_{\mathcal{I}})$

Proof.

$$I(i_x \leftrightarrow i_y | i_{\mathcal{I}}) \Rightarrow I(i_x \rightarrow i_y | i_{\mathcal{I}}) \tag{1}$$

$$\& I(i_y \rightarrow i_x | i_{\mathcal{I}}) \tag{2}$$

$$I(i_y \leftrightarrow i_z | i_{\mathcal{I}}) \Rightarrow I(i_y \rightarrow i_z | i_{\mathcal{I}}) \tag{3}$$

$$\& I(i_z \rightarrow i_y | i_{\mathcal{I}}) \tag{4}$$

$$(1) \& (3) \Rightarrow I(i_x \rightarrow i_z | i_{\mathcal{I}}) \tag{5}$$

$$(4) \& (2) \Rightarrow I(i_z \rightarrow i_x | i_{\mathcal{I}}) \tag{6}$$

$$(5) \& (6) \Rightarrow I(i_x \leftrightarrow i_z | i_{\mathcal{I}})$$

□

B.2.3.3 Algorithm for identifying drug interchangeabilities

Here I present the full algorithm behind the drug interchangeability identification as introduced in section 5.2.3 in the main text. In comparison to the simplified version in the main text, this version explicitly states at which points we can use information to prune the interchangeability space explored in the algorithm.

B.2. MATHEMATICAL DETAILS ON DRUG TARGET IDENTIFICATION

```

for  $\mathcal{I}_1 \in \mathcal{P}_{maxcard}(\mathbb{I})$  do
  if  $\mathcal{I}_1 \in \text{blacklist}$  then
    continue
  end if
  if  $T(\mathcal{I}_1)$  then
    add trivial extensions of  $\mathcal{I}_1$  to blacklist
    for  $\mathcal{I}_2 \in \mathcal{P}_{maxcard}(\mathcal{I}_1)$  do
      for  $\mathcal{I}_3 \in \mathcal{P}_{maxcard}(\mathbb{I} \setminus \mathcal{I}_1)$  do
        if  $I(i_{\mathcal{I}_2} \leftrightarrow i_{\mathcal{I}_3} | i_{\mathcal{I}_1})$  or interchangeability can be inferred from transitivity then
          report interchangeability
          add  $\mathcal{I}_1 \setminus \mathcal{I}_2 + \mathcal{I}_3$  and its trivial extensions to blacklist
        end if
      end for
    end for
  end if
end for

```

The running time of this algorithm is reduced by the application of pruning rules and defining a maximal cardinality in the power sets of drugs tested in the treatments. It should be noted that I assume the power sets to be ordered by cardinality.

B.2.4 Extending my definition of network selectivity

B.2.4.1 Selectivity in higher order drug treatments

For most applications a treatment will have to involve a larger number of drugs applied in parallel. In order to quantify the selectivity on each target in such a treatment, I extend the aforementioned definitions of network selectivity from chapter 5. First, I define the conditional required concentration of a drug within a certain treatment θ as

$$\xi_j(\mathcal{X}_p^2, \theta) = \min_{I_j} : \mathcal{X}_p^2(I_j \cdot \vec{e}_j + \theta) < 1.$$

In this definition the vector θ should include non-zero values for all drugs applied in parallel, except for drug j . Second, using this definition the conditional network selectivity reads

$$selectivity_j(\theta) = \frac{\xi_j(\mathcal{X}_{host}^2, \theta)}{\xi_j(\mathcal{X}_{parasite}^2, \theta)}.$$

This conditional selectivity describes how much effort has to be put into the design of each actual drug candidate in the treatment. In general, treatments

B.2. MATHEMATICAL DETAILS ON DRUG TARGET IDENTIFICATION

having a high conditional selectivity along all their drugs are preferential to those having low values for some of them, as those drugs might become a bottleneck in later development stages.

B.2.4.2 Selectivity in between multiple models

A further complication may arise in cases in which we need to consider the effect of a treatment on more than two distinct models. This might be the case if several models are necessary to cover potential side-effect in a host. In such cases I propose to compare the selectivity between the parasitic pathway and all host models and regard the lowest of these selectivities as the limiting one. Following this approach a resulting treatment has a high probability of achieving efficacy and safety with respect to all known side-effects.

B.2.5 A mathematical model to relate necessary drug concentrations to probabilities of resistance development

Depending on the type of disease which is to be cured by a simulated treatment, the possibility of resistance development should be considered when targets are selected. Therefore, I will introduce a few considerations in the following subsection, in which I will link the probability of resistance development against a certain drug to the necessary inhibitor concentrations. For simplification, I will make a number of assumptions.

First, I assume that the treatment consists of a single drug targeting a single enzyme in a non-competitive manner. Second, this enzyme is involved in an essential metabolic pathway, whose malfunction will lead to a serious impairment of the treated cell, e.g. a parasite. Third, I assume a strong correlation between the number of gene copies and the enzyme levels in the cell, i.e. if a drug resistance occurs in one gene, half of the enzyme is still susceptible to the treatment. Finally, it is supposed that the resistance to a certain drug can be described by a factor α , which signifies how much of the enzyme is insusceptible to the drug. As previously described, the reaction velocity of a reaction targeted with a non-competitive inhibitor can be modelled using the formula

$$V' = V \frac{1}{1 + \frac{I}{K_I}},$$

where V and V' are the reaction velocities before and after treatment, I is the concentration of the non-competitive inhibitor, and K_I is its binding

B.2. MATHEMATICAL DETAILS ON DRUG TARGET IDENTIFICATION

constant, then the reaction velocity after treatment given that a resistance has been developed can be described by

$$\begin{aligned} V' &= \alpha V + (1 - \alpha)V \frac{1}{1 + \frac{I}{K_I}} \\ &= V \frac{1 + \alpha \frac{I}{K_I}}{1 + \frac{I}{K_I}}. \end{aligned} \tag{B.4}$$

A value of $\alpha = 1$ would describe a full resistance to the treatment, while a complete susceptibility is modelled using $\alpha = 0$.

Given these assumptions a single, first mutation to a diploid organism, which is homozygote for the wild-type allele of the targeted enzyme throughout the population, can only affect one allele, and therefore render at most half of the available enzyme insensitive to the treatment, i.e. $\alpha \leq \frac{1}{2}$. One can argue now that if a drug only needs to inhibit a small fraction of the enzymes, e.g. 25 %, then a single mutation event, which can only lead to resistance in a single gene copy, will not be able to rescue the cell. At most half of the available enzyme will function properly, which is by assumption still lethal to the organism. After the mutation has occurred a higher drug dose might become necessary, but the treatment will still work rendering the mutation recessive. Opposed to that, if it is necessary to inhibit a large fraction of the enzyme, e.g. 75 %, for the drug to show an effect, a mutation is potentially able to lead to drug resistance by making 50 % of the enzyme insusceptible for the inhibitor. As such an effect cannot be accounted for by increasing the drug concentration, the mutation will appear to be dominant, because it rescues the organism already after the first mutation.

Depending on whether mutations in the targeted enzyme are preferably dominant or recessive, one can assume that the rates of resistance development vary as it is caused by either one or two mutation events. Thus, for treatments in which resistance is an issue one should select targets in which mutations are most likely to be recessive. In order to quantify the probability of a mutation in the targeted enzyme to be recessive, I will make the following considerations.

I assume that I am given an effective and safe inhibitor, which satisfies Equation 5.1, i.e. $\mathcal{X}^2 \left(\frac{I}{K_I} \right) \leq 1$. Given the essentiality of the pathway for the organism and the essentiality of the target for the pathway, the effect of the inhibitor on the network is exerted only via the inhibition of this single reaction. Thus, \mathcal{X}^2 is a function of the rate of the inhibited reaction V' , i.e. for the inhibitor $\mathcal{X}^2 \left(V' \left(\frac{I}{K_I} \right) \right) \leq 1$ holds. In terms of this function, the question of whether a certain mutation is recessive is equivalent to the

B.2. MATHEMATICAL DETAILS ON DRUG TARGET IDENTIFICATION

question of whether

$$\exists \frac{I'}{K_I} \geq 0 V' \left(\frac{I}{K_I}, \alpha = 0 \right) = V' \left(\frac{I'}{K_I}, \alpha \geq 0 \right)$$

holds. Inserting the definition from Equation B.4, we get

$$\begin{aligned} \exists \frac{I'}{K_I} \geq 0 V \frac{1}{1 + \frac{I}{K_I}} &= V \frac{1 + \alpha \frac{I'}{K_I}}{1 + \frac{I'}{K_I}} \\ \Leftrightarrow \exists \frac{I'}{K_I} \geq 0 \frac{I'}{K_I} &= \frac{-\frac{I}{K_I}}{\alpha \frac{I}{K_I} + \alpha - 1}, \end{aligned}$$

which holds if and only if

$$\frac{I}{K_I} < \frac{1}{\alpha} - 1. \quad (\text{B.5})$$

Using Equation B.5 and given knowledge of a distribution for α values, one can compute a probability distribution for the fact of whether a beneficial mutation that affects a targeted enzyme is recessive or dominant. In case one assumes a uniform distribution of α values in the range of 0 to $\frac{1}{2}$, the cumulative distribution of the probability of the mutation being recessive can be given by

$$\text{Pr} = \frac{2}{\frac{I}{K_I} + 1}.$$

Therefore, I conclude that those targets should be preferred, which have a lower $\frac{I}{K_I}$ value needed for an effective and safe treatment, as mutations affecting this enzyme will be less likely to be dominant and will therefore less likely manifest on a population level. Despite these clear and intuitive results, the assumptions made in the beginning are quite restrictive. Specific assumptions on the effect of the treatment, the target organism, and the way in which resistance arise have been made. As these assumptions cannot be assumed to be universal, e.g. resistances can arise through various mechanisms, not only mutations in the enzyme targeted by a drug, the results should be treated with caution. Nevertheless, they do indicate ideas on how resistance development could be slowed down and they support my claim that targets having the lowest necessary $\frac{I}{K_I}$ values are to be preferred.

B.3. MATHEMATICAL DETAILS OF SYNERGISM ANALYSIS

B.3 Mathematical details of synergism analysis

B.3.1 Rewriting Bliss drug interaction models

Bliss independence The formula for the Bliss model in its original form is $fu_{12} = fu_1fu_2$, meaning the the fraction unaffected (e.g. the fraction of bacteria surviving a treatment) after combined medication (fu_{12}) is the product of the fractions after single medication ($fu_{1/2}$). Transferred to a setting in which we observe the concentration of a single species x , which is elevated in the pathological state, this independence can be written as

$$\begin{aligned} \frac{x_{12}}{x_0} &= \frac{x_1}{x_0} \cdot \frac{x_2}{x_0} && \Leftrightarrow \\ x_{12} &= \frac{x_1x_2}{x_0}, \end{aligned}$$

where x_0 is the untreated concentration of x after a given time, $x_{1/2}$ are the concentrations after treatment with a single drug, and x_{12} is the concentration after dual treatment.

Bliss boosting A simple version of the Bliss boosting model from Lehar *et al.* [Lehár et al., 2007] can be written in terms of the “affected fraction” $fa_{1/2/12} = 1 - fa_{1/2/12}$:

$$\begin{aligned} 1 - fa_{12} &= (1 - fa_1)(1 - fa_2) && \Leftrightarrow \\ fa_{12} &= fa_1 + fa_2 - fa_1fa_2. \end{aligned}$$

A boosting factor α is now introduced in such a way that a value $\alpha = 1$ represents independence, larger values represent antagonisms, and smaller values synergisms. The definition is not completely coherent with the original formula of Lehar *et al.* . In our formulation we leave out the explicit mentioning of enzyme concentrations (because they do not necessarily appear in the model). Therefore, the Bliss boosting model reduces to

$$\begin{aligned} fa_{12} &= fa_1 + fa_2 - \alpha fa_1fa_2 && \Leftrightarrow \\ 1 - \frac{x_{12}}{x_0} &= \left(1 - \frac{x_1}{x_0}\right) + \left(1 - \frac{x_2}{x_0}\right) - \alpha \left(1 - \frac{x_1}{x_0}\right) \cdot \left(1 - \frac{x_2}{x_0}\right) && \Leftrightarrow \\ x_{12} &= x_1 + x_2 - x_0 + \alpha * \left(x_0 - x_1 - x_2 + \frac{x_1x_2}{x_0}\right) \end{aligned}$$

B.3. MATHEMATICAL DETAILS OF SYNERGISM ANALYSIS

Bliss independence of χ^2 values In the case in which the objective function behind our drug target identification optimisation only consists of a single concentration the independence models are comparably simple. For examples in which we are using a χ^2 value, the situation is getting more difficult. Let us assume that we want to fit 2 simulated timepoints (x and y) to a value of 0 with a standard deviation of 1. In this simple case, the objective function reduces to $\chi_{1/2/12}^2 = x_{1/2/12}^2 + y_{1/2/12}^2$.

Assuming Bliss independence between two inhibitors the formula for the objective function under dual inhibition reads

$$\chi_{12}^2 = \left(\frac{x_1 x_2}{x_0}\right)^2 + \left(\frac{y_1 y_2}{y_0}\right)^2,$$

which cannot generally be rewritten in terms of $\chi_{0/1/2}^2$. There are only two conditions under which this is possible.

- First, if the objective function consists of a single timepoint, we have $\chi_{1/2}^2 = x_{1/2}^2$ and $\chi_{12}^2 = \frac{x_1^2 x_2^2}{x_0^2} = \frac{\chi_1^2 \chi_2^2}{\chi_0^2}$. This formula can be generalized for the case in which n inhibitors affect one single timepoints:

$$\chi_{1..n}^2 = \prod_i^n \chi_i^2 \cdot \chi_0^{2-2n}.$$

- Second, if each of the two inhibitors affects only one timepoint, we have w.l.o.g. $x_2 = x_0$ and $y_1 = y_0$ and our formula reduces to $\chi_{12}^2 = x_1^2 + y_2^2 = \chi_1^2 + \chi_2^2 - \chi_0^2$. This formula can also be generalized for n timepoints which are all individually affected by n inhibitors. In the general case it reads

$$\chi_{1..n}^2 = \sum_i^n \chi_i^2 - (n-1)\chi_0^2.$$

Both of these formulas can be extended for cases in which the χ^2 value contains summands that are affected by neither drug, e.g. $\chi_{1/2/12}^2 = x_{1/2/12}^2 + c$ with the x timepoint being affected by 2 inhibitors and multiple not affected timepoints in the residual sum c . With this extension the independence model for the dual inhibition reads

$$\chi_{12}^2 = \frac{(\chi_1^2 - c)(\chi_0^2 - c)}{\chi_0^2 - c} + c.$$

B.3. MATHEMATICAL DETAILS OF SYNERGISM ANALYSIS

Appendix C

Models and objective functions for the target identification

C.1 Linear chain



Supplementary Figure C.1: Structure of the linear model as used in Gerber *et al.* [Gerber *et al.*, 2008] visualised using biographer [Handorf *et al.*, 2012].

The chain model investigated in this work implements a linear reaction network as shown in Figure C.1. Reaction velocities are described by reversible Michaelis-Menten-kinetics with parameters $V_{\max}^f = K_{mS} = K_{mP} = 1$, $V_{\max}^r = 0.2$ and species concentrations $S_1 = 1$ and $S_6 = 0$ are fixed while the others are variables of the model. To distinguish, which parameters influence the result of the tested inhibitions, I have selected different parameter sets. These will be introduced together with the specific results later in this section.

C.1.1 Potency of different inhibition types

In order to explain some of the results introduced in the main text, I would like to investigate how inhibitors with different modes-of-action influence the velocity of reversible Michaelis-Menten kinetics. For all considerations I will assume that the reaction runs in forward direction, i.e. $V_{\max}^f \frac{S}{K_{mS}} - V_{\max}^r \frac{P}{K_{mP}} > 0$, and that the parameters and concentrations are non-negative but are not subject to any further constraints.

C.1. LINEAR CHAIN

Comparing competitive and uncompetitive inhibition In the following, I will try to deduce conditions under which an uncompetitive inhibitor has a larger effect on a reaction velocity than a competitive inhibitor at a comparable effective concentration.

$$\begin{aligned}
 V(I_c) > V(I_u) & \Leftrightarrow \\
 \frac{V_{\max}^f \frac{S}{K_{mS}} - V_{\max}^r \frac{P}{K_{mP}}}{\frac{S}{K_{mS}} + \frac{P}{K_{mP}} + \left(1 + \frac{I_c}{K_{I_c}}\right)} > \frac{V_{\max}^f \frac{S}{K_{mS}} - V_{\max}^r \frac{P}{K_{mP}}}{\left(1 + \frac{I_u}{K_{I_u}}\right) \cdot \left(\frac{S}{K_{mS}} + \frac{P}{K_{mP}}\right) + 1} & \Leftrightarrow \\
 1 + \frac{S}{K_{mS}} + \frac{P}{K_{mP}} + \frac{I_c}{K_{I_c}} < 1 + \frac{S}{K_{mS}} + \frac{P}{K_{mP}} + \frac{I_u}{K_{I_u}} \frac{S}{K_{mS}} + \frac{I_u}{K_{I_u}} \frac{P}{K_{mP}} & \Leftrightarrow \\
 \frac{I_c}{K_{I_c}} < \frac{I_u}{K_{I_u}} \frac{S}{K_{mS}} + \frac{I_u}{K_{I_u}} \frac{P}{K_{mP}}. &
 \end{aligned}$$

Given identical effective inhibitor concentrations, i.e. $\frac{I_c}{K_{I_c}} = \frac{I_u}{K_{I_u}}$, the uncompetitive inhibition is stronger than the competitive inhibition if and only if

$$1 < \frac{S}{K_{mS}} + \frac{P}{K_{mP}}. \quad (\text{C.1})$$

This equation defines a mathematical condition under which an uncompetitive inhibitor should be preferred over a competitive one. It can intuitively be understood via the idea that the inhibitor works best if it is binding the most abundant form of the enzyme. E.g. given $S \ll K_{mS}$ and $P \ll K_{mP}$ the binding of substrate and product is weak, the enzyme is largely unbound, and the competitive inhibitor is more effective than the uncompetitive one.

Comparing competitive and non-competitive inhibition Starting from the same assumptions I will now compare the effects of a competitive and a non-competitive inhibitor on the reaction velocity of a reversible enzymatic reaction.

$$\begin{aligned}
 V(I_c) \geq V(I_n) & \Leftrightarrow \\
 \frac{V_{\max}^f \frac{S}{K_{mS}} - V_{\max}^r \frac{P}{K_{mP}}}{\frac{S}{K_{mS}} + \frac{P}{K_{mP}} + \left(1 + \frac{I_c}{K_{I_c}}\right)} \geq \frac{V_{\max}^f \frac{S}{K_{mS}} - V_{\max}^r \frac{P}{K_{mP}}}{\frac{S}{K_{mS}} + \frac{P}{K_{mP}} + 1} \cdot \frac{1}{1 + \frac{I_n}{K_{I_n}}} & \Leftrightarrow \\
 1 + \frac{S}{K_{mS}} + \frac{P}{K_{mP}} + \frac{I_c}{K_{I_c}} \leq \left(\frac{S}{K_{mS}} + \frac{P}{K_{mP}} + 1\right) \cdot \left(1 + \frac{I_n}{K_{I_n}}\right) & \Leftrightarrow \\
 \frac{I_c}{K_{I_c}} \leq \frac{I_n}{K_{I_n}} + \frac{S}{K_{mS}} \frac{I_n}{K_{I_n}} + \frac{P}{K_{mP}} \frac{I_n}{K_{I_n}} &
 \end{aligned}$$

Assuming that we are given comparable effective inhibitor concentrations and that all parameters are non-negative, a non-competitive inhibition will always be at least as good as a competitive inhibition. Furthermore, equivalence of both modes is only achieved iff $S = P = 0$, otherwise a non-competitive inhibitor will be better.

Comparing uncompetitive and non-competitive inhibition The last pair of modes-of-action, that is compared, is uncompetitive and non-competitive inhibition.

$$\begin{aligned}
 V(I_u) > V(I_n) & \Leftrightarrow \\
 \frac{V_{\max}^f \frac{S}{K_{mS}} - V_{\max}^r \frac{P}{K_{mP}}}{\left(1 + \frac{I_u}{K_{I_u}}\right) \cdot \left(\frac{S}{K_{mS}} + \frac{P}{K_{mP}}\right) + 1} > \frac{V_{\max}^f \frac{S}{K_{mS}} - V_{\max}^r \frac{P}{K_{mP}}}{\frac{S}{K_{mS}} + \frac{P}{K_{mP}} + 1} \cdot \frac{1}{1 + \frac{I_n}{K_{I_n}}} & \Leftrightarrow \\
 \left(1 + \frac{I_u}{K_{I_u}}\right) \cdot \left(\frac{S}{K_{mS}} + \frac{P}{K_{mP}}\right) + 1 < \left(\frac{S}{K_{mS}} + \frac{P}{K_{mP}} + 1\right) \cdot \left(1 + \frac{I_n}{K_{I_n}}\right) & \Leftrightarrow \\
 \frac{I_u}{K_{I_u}} \left(\frac{S}{K_{mS}} + \frac{P}{K_{mP}}\right) < \frac{I_n}{K_{I_n}} \left(1 + \frac{S}{K_{mS}} + \frac{P}{K_{mP}}\right) &
 \end{aligned}$$

Given non-negative parameters, a non-competitive inhibitor will always have a larger effect on a reaction velocity than an uncompetitive inhibitor at the same effective concentration. Furthermore, it can be seen that the advantage of the non-competitive inhibitor gets smaller for increasing effective substrate and product concentrations.

C.1.2 Setting up the objective function

In their article, Gerber *et al.* observed the flux through the linear chain under different inhibitions. While the flux through a reaction is a variable that can easily be analysed in MCA, dynamic simulations only keep direct track of the variables, i.e. the chemical species, of the model. Therefore, fluxes can only be determined by evaluating the reaction kinetics (including all inhibitors) given the current concentrations. For the standard parametrisation of the model (as presented in the main text), the objective function judging the flux is given by

$$\begin{aligned}
 \mathcal{X}^2 &= \left(\frac{V_5}{V_5^{\text{ss}}/2}\right)^2 && \text{with} \\
 v &= \frac{V_{\max}^f \frac{S_5}{K_{mS}} - V_{\max}^r \frac{S_6}{K_{mP}}}{\left(1 + \frac{I_u}{K_{I_u}}\right) \cdot \left(\frac{S_5}{K_{mS}} + \frac{S_6}{K_{mP}}\right) + \left(1 + \frac{I_c}{K_{I_c}}\right)} \cdot \frac{1}{1 + \frac{I_n}{K_{I_n}}}, && \text{(C.2)}
 \end{aligned}$$

I_c , I_u , and I_n being the concentration of the competitive, the uncompetitive, and the non-competitive inhibitor, and $V_5^{\text{ss}} \approx 0.272$ being the steady state flux through reaction 5 without inhibitions. The objective function is constructed as such that it will drop below 1 if the flux is reduced to less than half its steady state value, i.e. $V_5 < \frac{V_5^{\text{ss}}}{2}$.

Gerber *et al.* argue, that the potency of inhibitors along the chain has to be judged by flux control coefficients and not by concentration control coefficients, e.g. the derivative of the steady state concentration of S_5 with respect

C.1. LINEAR CHAIN

to the velocity of an inhibited reaction. Analyses observing the concentration control coefficients would be misleading as the coefficients switch signs when the inhibition targets a reaction “behind” the considered substance. For my approach, however, fluxes are less easily handable, as experimental data for the “healthy” state is harder to determine experimentally than metabolite concentrations. Therefore, I want to investigate conditions under which observing a metabolite concentration provides comparable results to settings in which the change in a flux is observed.

I propose the following three conditions under which the concentration of a certain metabolite can be representative of the flux: First, a reaction r has to be selected, whose velocity is linearly dependent on the flux through the network because of stoichiometric constraints, e.g. it is one of the reactions in a linear pathway. Second, the selected metabolite s is the only variable in the kinetics of reaction r , i.e. the only substrate or product with no further modifiers present. Third, the velocity of r is monotonic with respect to s , i.e. no substrate inhibition is assumed.

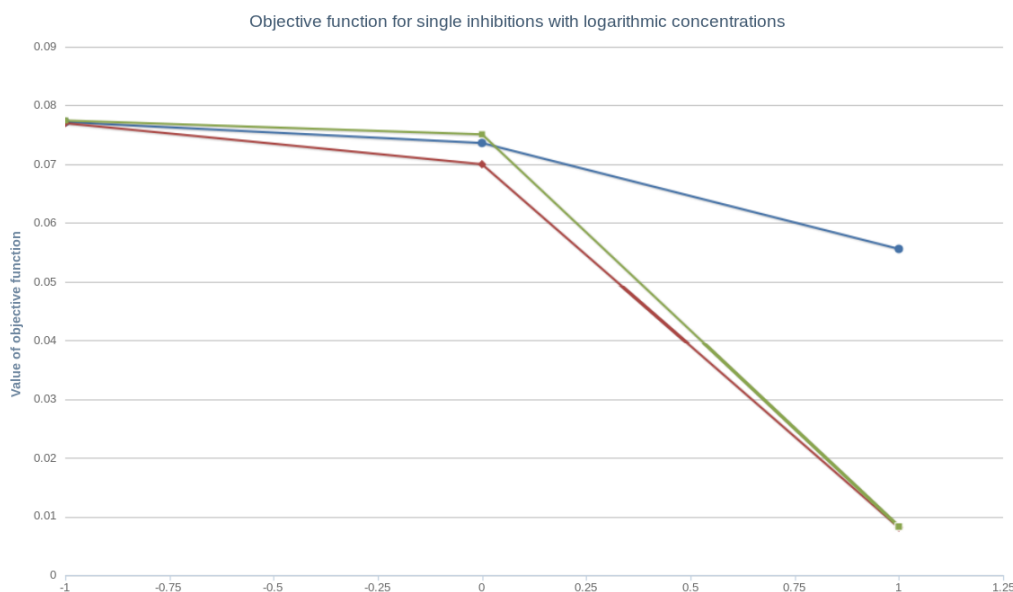
Under these conditions s can be representative for the flux through the network, given that two points are kept in mind. First, the relationship between the flux and s is non-linear. Second, inhibition of reactions which appear after s in the pathway have to be neglected [Gerber et al., 2008]. This is for example underlined by Eq. C.3. It shows that an inhibition of the last reaction changes the relation between s and the flux through the reaction.

In the example of the linear chain one can in principle observe the concentration of substance S_5 instead of the flux. The resulting objective function then becomes

$$\begin{aligned} \chi^2 &= \left(\frac{S_5}{S_5^{\text{ss}}/2} \right)^2 && \text{with} \\ S_5^{\text{ss}} &= 0.3862, \end{aligned} \tag{C.3}$$

with the advantage that the “healthy” concentration of S_5 can be determined more easily experimentally. When investigating inhibitions of the first four reactions, the effects on S_5 and V_5 are comparable and easily explainable. However, when observing the effect of different inhibitor modes-of-action on reaction 5 an interesting behaviour becomes apparent (see Figure C.2): For low inhibitor concentrations the competitive inhibitor is stronger than the uncompetitive one. This can be explained by the relation shown in Eq. C.1. For inhibitor concentrations above $I_c, I_u \approx 1.75$ the steady state concentration of S_5 rises above 1 and Eq. C.1 switches from being false to true. An interesting follow-up question is why does this effect only show up for reaction 5. This can be explained by the fact that the product of reaction

C.1. LINEAR CHAIN



Supplementary Figure C.2: Abnormality concerning different inhibitions of the last reaction in the chain. The Figure shows the effects on the flux (see Eq.C.3) through the chain after inhibition with different modes of action (red:noncompetitive, blue:competitive, green:uncompetitive) at different concentrations (x-axis). At other positions in the network competitive inhibition is the weakest mode-of-action at all concentrations, which is not the case at this reaction.

5 is kept constant at a low concentration of 0. For higher concentrations, which would change the outcome of inequality Eq. C.1 this behaviour would not be present. Therefore, I conclude that investigating reactions at the “border” of the network should be taken with a pinch of salt. Thus, when investigating the last metabolite in a chain instead of the flux (neglecting inhibitions of reactions behind it), it does not mean one necessarily loses relevant information.

C.1.3 Results

I have conducted different analyses on different parametrisations of the Gerber model. In the following, the various results and their underlying parameter values and objective functions are introduced.

In Figure C.3, I have investigated a version of the model in which the equilibrium constants of all reactions are set to 1. For this purpose I have chosen the parameter set $V_{\max}^f = V_{\max}^r = K_{mS} = K_{mP} = 1$ and investigated

C.1. LINEAR CHAIN

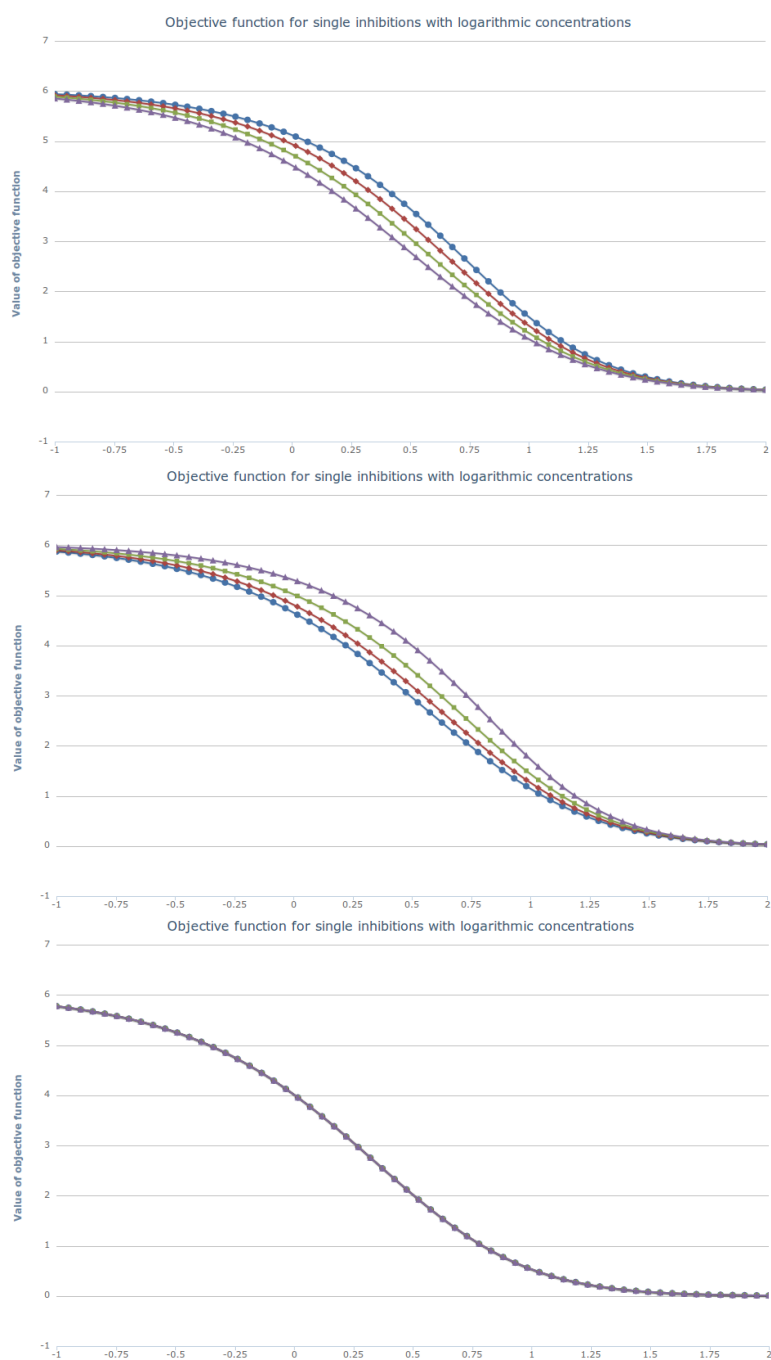
changes in the objective function $\mathcal{X}^2 = \left(\frac{S_5}{0.05}\right)^2$. Using this objective function instead of one focussing on the flux leaves one with results mostly comparable to those of Gerber *et al.* .

For Figure C.4, I have changed the equilibrium constants of all reactions to 100 using the parameter set $V_{\max}^f = K_{mP} = 10, V_{\max}^r = K_{mS} = 1$ and investigated changes in the objective function $\mathcal{X}^2 = \left(\frac{S_5}{0.446}\right)^2$.

For Figure C.5, I have changed the parameter set to $V_{\max}^f = V_{\max}^r = K_{mS} = K_{mP} = 100$ and investigated changes in the objective function $\mathcal{X}^2 = \left(\frac{S_5}{0.0992}\right)^2$.

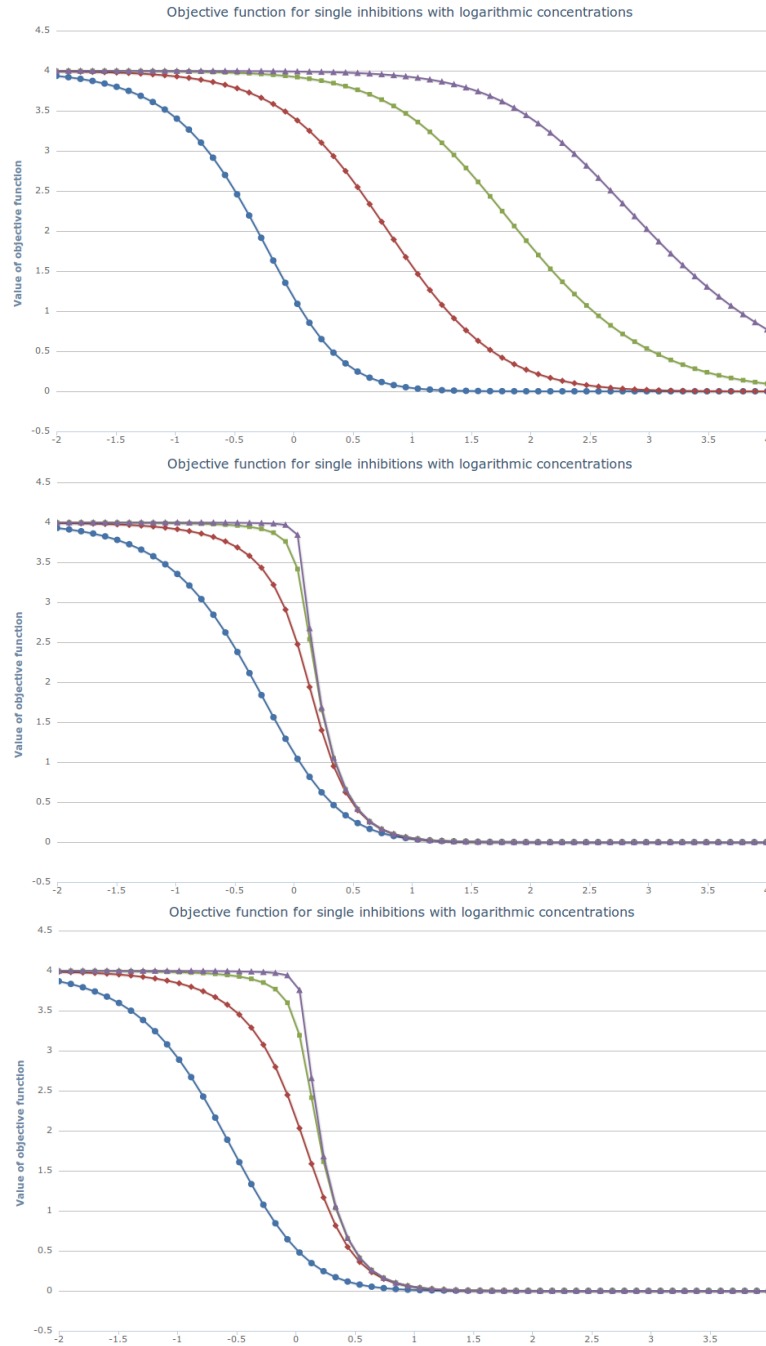
The results of these simulations support the main conclusions drawn in section 5.3.2.2 in the main text.

C.1. LINEAR CHAIN



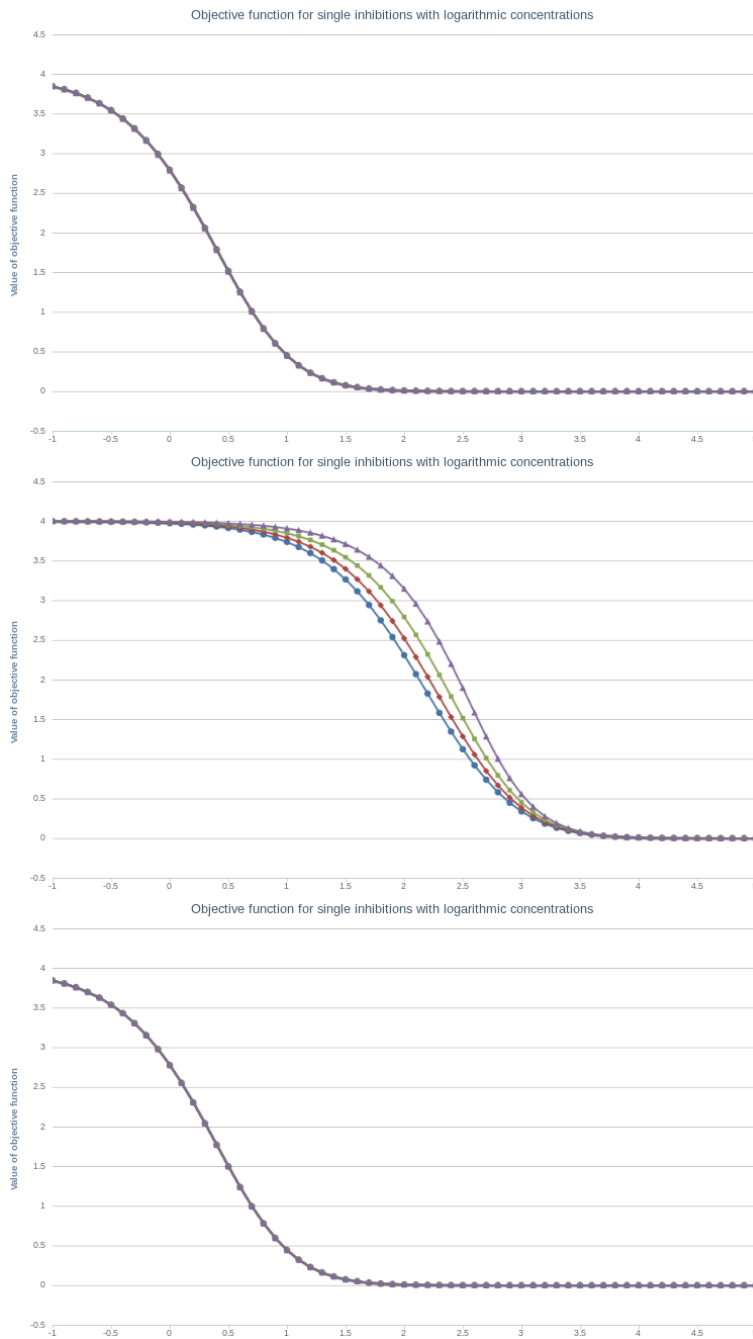
Supplementary Figure C.3: Effects of different inhibitors on a linear chain in which all parameters have been set to 1. The three graphs show the effects of competitive (top), uncompetitive (middle), and non-competitive (bottom) inhibitors on the objective function (y-axis) for different positions (reaction 1: blue, 2: red, 3: green, 4: violet) in varying effective concentrations (x-axis in \log_{10} scale).

C.1. LINEAR CHAIN



Supplementary Figure C.4: Same analysis as performed in Figure C.3, except for the fact that the equilibrium constants have been changed from 1 to 100.

C.1. LINEAR CHAIN



Supplementary Figure C.5: Same analysis as performed in Figure C.3. All parameters have been set to 100 to demonstrate general rules in the inhibitor's response to changes in the parameter set.

C.2 Glycolysis in *Trypanosoma brucei*

Model describes the glycolysis in the bloodstream form of *Trypanosoma brucei*, the pathogen causing sleeping sickness. In the scope of this work I will compare results from two different models, the model of Albert *et al.* [Albert *et al.*, 2005] and the model of Achcar *et al.* [Achcar *et al.*, 2012]. The latter of the models is an updated version of the Albert model, it however fails to reproduce one result of Visser [Visser, 1981]. In Visser's experiments the total glucose consumption is not altered in between aerobic and anaerobic conditions, i.e. half of the flux directed towards pyruvate is redirected towards glycerol under anaerobic conditions. Because this result has not been considered by Achcar *et al.*, single inhibitions of the glycerol 3 phosphate oxidase reaction (as compared to inhibitions accompanied by an increase in external glycerol) appear to be far more potent than they might be [Nok, 2002, Minagawa *et al.*, 1997]. Therefore, I will consider both models instead of just the updated one.

C.2.1 Setting up the objective function

As the objective function judging the performance of the glycolysis, I have chosen the flux through the pyruvate transporter, which should be inhibited by 50% in order to kill the pathogen [Bakker *et al.*, 1999]. Therefore, the objective function for the Albert model is

$$\mathcal{X}^2 = \left(\frac{\frac{200 \cdot \text{species1}}{1.96 + \text{species1}} \cdot \frac{1}{1 + \frac{\text{nonc.vPT}}{1}} \cdot \left(1 + \frac{\text{a.vPT}}{1}\right)}{87.25} \right)^2 \quad (\text{C.4})$$

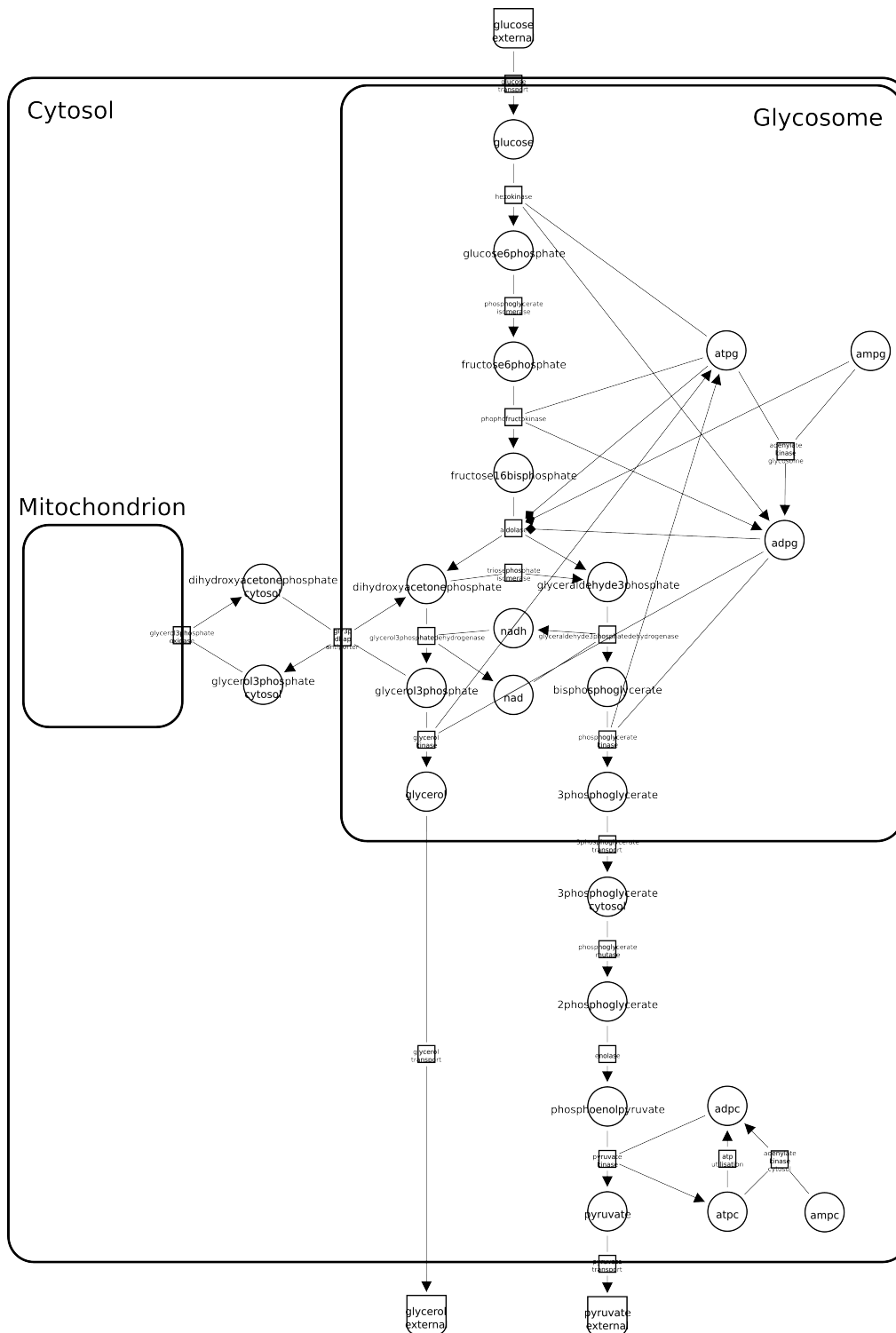
and the objective function for the Achcar model has been set to

$$\mathcal{X}^2 = \left(\frac{\frac{200 \cdot \text{Pyr.c}}{1.96 + \text{Pyr.c}} \cdot \frac{1}{1 + \frac{\text{nonc.vPyrT.c}}{1}} \cdot \left(1 + \frac{\text{a.vPyrT.c}}{1}\right)}{89.8} \right)^2. \quad (\text{C.5})$$

C.2.2 Results

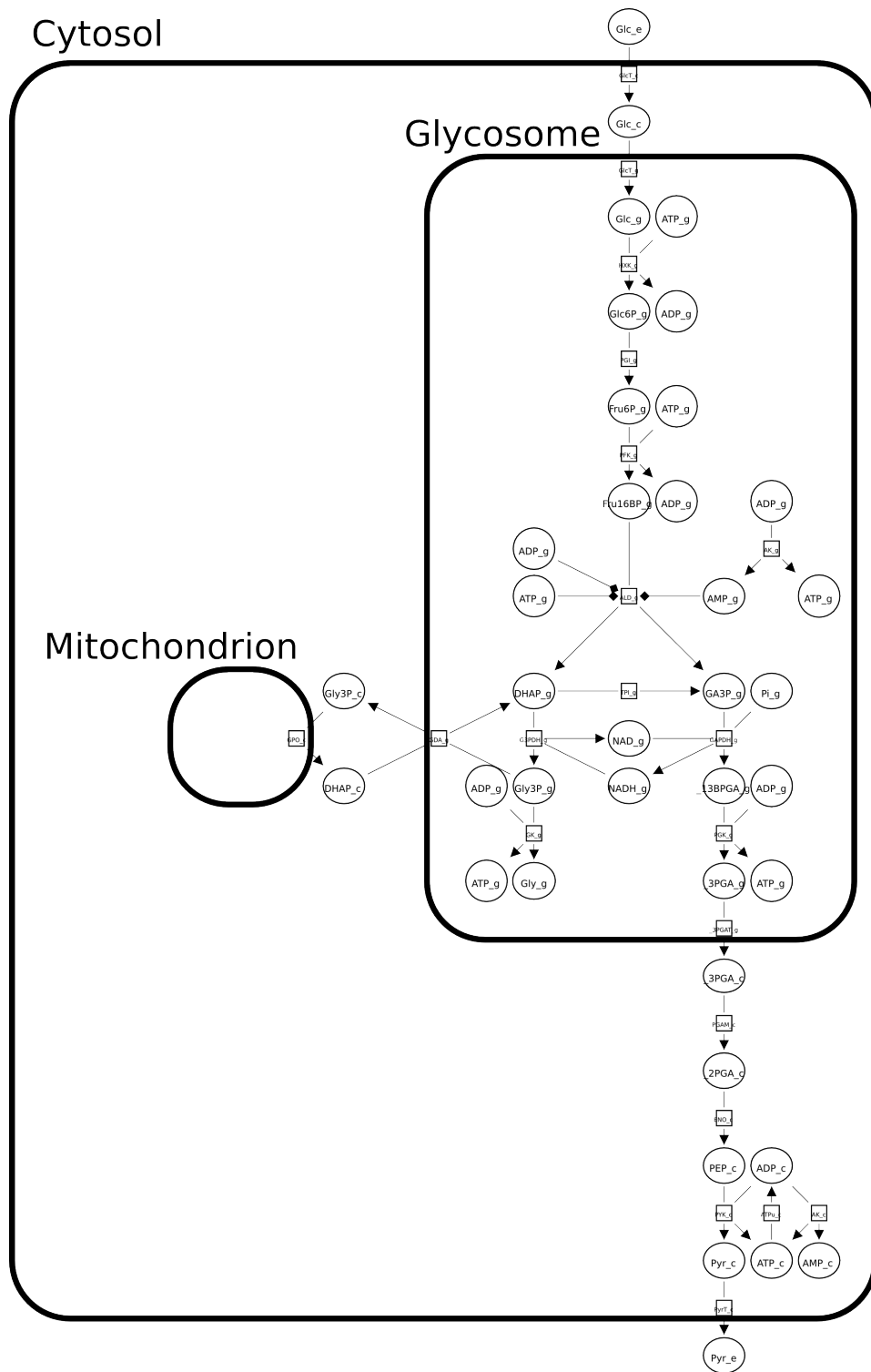
Results of Table C.1 are largely in agreement with results produced using a similar method [Schulz *et al.*, 2009]. Small differences result from a slightly different objective function (comparing flux through upper glycolysis with flux through lower glycolysis), but they are in the order of magnitude of the variance resulting from simulating the ODE system using different solvers with different parameters.

C.2. GLYCOLYSIS IN *TRYPANOSOMA BRUCEI*



Supplementary Figure C.6: Structure of the glycolysis in *Trypanosoma brucei* as described by Albert *et al.* [Albert *et al.*, 2005].

C.2. GLYCOLYSIS IN *TRYPANOSOMA BRUCEI*



Supplementary Figure C.7: Structure of the glycolysis in *Trypanosoma brucei* as described by Achcar *et al.* [Achcar *et al.*, 2012].

C.2. GLYCOLYSIS IN *TRYPANOSOMA BRUCEI*

Supplementary Table C.1: Inhibitor concentrations necessary for an effective treatment. (c: competitive inhibition, k: competitive inhibition on a cofactor binding site, n: non-competitive inhibition, u: uncompetitive inhibition, ALD: aldolase, AU: ATP utilisation, ENO: enolase, GAPDH: glyceraldehyde-3-phosphate dehydrogenase, GlcTg: glucose transport into glycosome, GPDH: glycerol-3-phosphate dehydrogenase, THT: trypanosoma hexose transport, HK: hexokinase, PGI: glucose-6-phosphate isomerase, PGK: phosphoglycerate kinase, PGM: phosphoglycerate mutase, PK: pyruvate kinase, PT: pyruvate transport, TPI: triosephosphate isomerase)

| Target | MOA | Inhibitor concentration | |
|--------|-----|-------------------------|-------------|
| | | Albert 2005 | Achcar 2012 |
| THT | n | 1.07 | 1.09 |
| THT | u | 1.28 | 1.29 |
| THT | n | 1.29 | 1.31 |
| PGM | n | 1.58 | 1.59 |
| PGM | u | 1.58 | |
| GAPDH | n | 2.51 | 2.82 |
| GPDH | n | 3.37 | 3.48 |
| GPDH | u | 3.44 | 3.56 |
| GAPDH | u | 4.71 | 4.86 |
| GAPDH | c | 5.48 | 6.67 |
| ENO | n | 5.86 | 5.95 |
| ENO | u | 5.86 | |
| THT | c | 6.46 | 6.53 |
| ALD | u | | 7.19 |
| ALD | n | 7.89 | 8.14 |
| PK | n | 9.78 | 9.84 |
| PK | u | | 9.84 |
| PGK | n | 24.9 | 25.1 |
| PGK | u | 25.1 | 25.3 |
| PGI | n | 27.7 | 28.1 |
| TPI | n | 27.9 | 28.5 |
| PGI | u | 27.9 | |
| GAPDH | k | 28 | 31.2 |
| TPI | u | 29.9 | |
| HK | n | 37.5 | 37.9 |
| HK | u | 41.7 | 42.2 |
| PFK | u | 138 | 18.4 |
| GPDH | c | 174 | 172 |
| PFK | n | 190 | 14.8 |
| PK | k | | 122 |
| GPDH | k | 204 | 226 |
| AU | n | 209 | 212 |
| PGK | k | 273 | 298 |
| HK | c | 372 | 378 |
| TPI | c | 481 | |
| PFK | c | 1610 | 177 |
| PFK | k | | 946 |
| PK | c | | 1280 |
| AU | a | | 1500 |
| HK | k | 1970 | 1980 |
| PGK | c | 2780 | 3420 |
| PGI | c | 2990 | |
| GlcTg | n | | 5160 |
| ALD | c | | 9150 |
| 3PGAT | n | | 11600 |
| PGM | c | 14700 | |
| ENO | c | 25100 | |

C.3. GLYCOLYSIS IN HUMAN ERYTHROCYTES

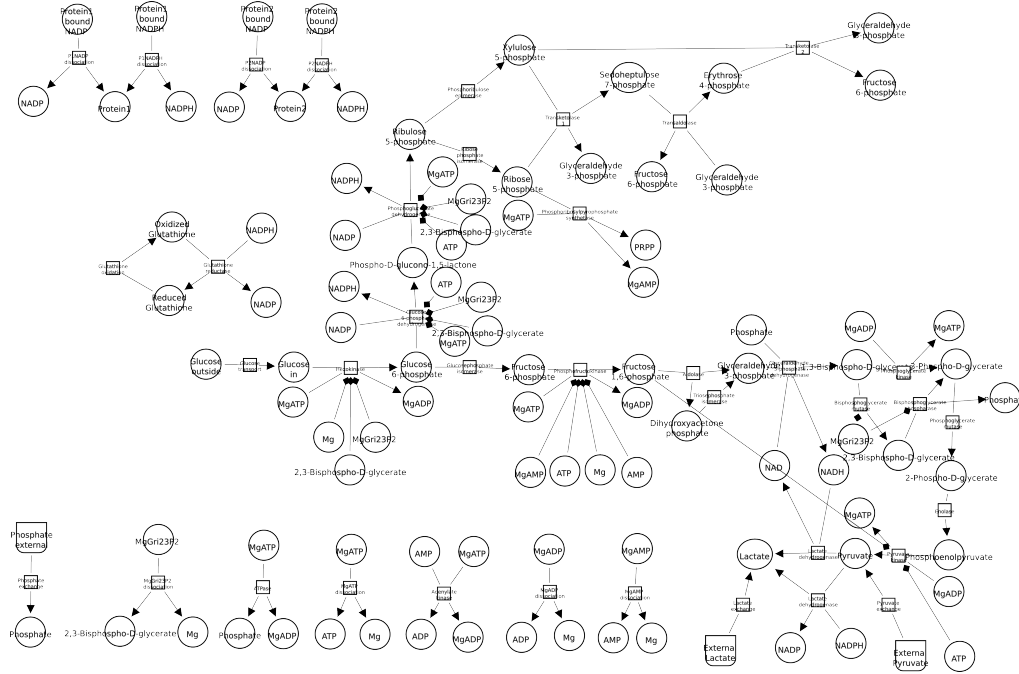
C.3 Glycolysis in human erythrocytes

| Model (click for information) | BioModel | Similarity | | Overlap | |
|-------------------------------------------------------------|---------------------------------|------------|---------|---------|----------|
| | | score | p-value | score | p-value |
| Albert2005_Glycolysis | BIOMD0000000211 | 1.000 | <=1e-3 | 64 | -4.4e-16 |
| Bakker2001_Glycolysis | BIOMD0000000071 | 0.692 | <=1e-3 | 42 | -4.4e-16 |
| Teusink2000_Glycolysis | BIOMD0000000064 | 0.522 | <=1e-3 | 36 | -4.4e-16 |
| Pritchard2002_glycolysis | BIOMD0000000172 | 0.499 | <=1e-3 | 32 | -4.4e-16 |
| Ralser2007_Carbohydrate_Rerouting_ROS | BIOMD0000000247 | 0.499 | <=1e-3 | 38 | -4.4e-16 |
| Conant2007_WGD_glycolysis_2A3AB | BIOMD0000000176 | 0.491 | <=1e-3 | 31 | -4.4e-16 |
| Hynne2001_Glycolysis | BIOMD0000000061 | 0.474 | <=1e-3 | 31 | -4.4e-16 |
| Holzhutter2004_Erythrocyte_Metabolism | BIOMD0000000070 | 0.471 | <=1e-3 | 41 | -4.4e-16 |
| Conant2007_glycolysis_2C | BIOMD0000000177 | 0.467 | <=1e-3 | 31 | -4.4e-16 |
| Teusink1998_Glycolysis_TurboDesign | BIOMD0000000253 | 0.465 | <=1e-3 | 20 | 1.7e-15 |
| Galazzo1990_FermentationPathwayKinetics | BIOMD0000000063 | 0.453 | <=1e-3 | 24 | -4.4e-16 |
| Nielsen1998_Glycolysis | BIOMD0000000042 | 0.433 | <=1e-3 | 29 | -4.4e-16 |
| Chassagnole2002_Carbon_Metabolism | BIOMD0000000051 | 0.398 | <=1e-3 | 33 | -4.4e-16 |
| Chance1960_Glycolysis_Respiration | BIOMD0000000281 | 0.349 | <=1e-3 | 18 | 4.3e-13 |
| Westermarck2003_Pancreatic_GlycOsc_extended | BIOMD0000000236 | 0.335 | <=1e-3 | 10 | 2.9e-05 |
| Wolf2000_Glycolytic_Oscillations | BIOMD0000000206 | 0.327 | <=1e-3 | 15 | 7.6e-10 |
| Poolman2004_CalvinCycle | BIOMD0000000013 | 0.279 | <=1e-3 | 18 | 4.3e-13 |
| Rohwer2001_Sucrose | BIOMD0000000023 | 0.176 | <=1e-3 | 5 | 5.4e-02 |
| Jiang2007_GSISsystem_PancreaticBetaCells | BIOMD0000000239 | 0.161 | <=1e-3 | 10 | 2.9e-05 |
| Westermarck2003_Pancreatic_GlycOsc_basic | BIOMD0000000225 | 0.155 | <=1e-3 | 3 | 3.4e-01 |
| Cronwright2002_Glycerol_Synthesis | BIOMD0000000076 | 0.130 | <=1e-3 | 3 | 3.4e-01 |
| Voit2003_Trehalose_Cycle | BIOMD0000000266 | 0.123 | 2.0e-03 | 2 | 6.2e-01 |
| Valero2006_Adenine_TernaryCycle | BIOMD0000000231 | 0.119 | 5.0e-03 | 3 | 3.4e-01 |
| Tyson2003_Mutual_Activation | BIOMD0000000311 | 0.112 | 1.2e-02 | 0 | 1.0e+00 |
| Rovers1995_Photosynthetic_Oscillations | BIOMD0000000292 | 0.111 | 1.2e-02 | 4 | 1.5e-01 |

Supplementary Figure C.8: Results of the model retrieval starting from the Albert model of the *Trypanosoma* glycolysis [Albert et al., 2005]. Most of the depicted models describe glycolysis in *Saccharomyces cerevisiae*, except for Holzhütter: human erythrocytes, Chassagnole: *Escherichia coli*, Chance: human tumor cells, Westermarck and Jiang: human pancreatic beta cell, and Poolman: *Nicotiana tabacum*. From the models describing glycolysis in human cells the Holzhütter model has the largest overlap with the Albert model.

In order to develop a treatment working *in vivo*, one not only needs to consider a drug's efficacy but also the potential side effects this treatment would have in a human. Given the case that a parasite infection should be treated, one class of potential side effects can be predicted by computing the effect of a treatment on the investigated pathway in the human host. Models being able to serve this purpose can be identified by using our model retrieval website. The results of the model retrieval can be found in Figure C.8.

C.3. GLYCOLYSIS IN HUMAN ERYTHROCYTES



Supplementary Figure C.9: Structure of the energy and redox metabolism in human erythrocytes as described by Holzhütter [Holzhütter, 2004].

C.3.1 Setting up the objective function

As a representative of the flux through the lower glycolysis I observed the flux through the enolase reaction, which is supposed to be inhibited at most by 5%. Using the objective function

$$\chi^2 = \left(\frac{1500 \frac{Gri2P - PEP}{1 + Gri2P + PEP}}{2.805} \right)^2$$

treatments reducing the flux by more than 5% will have an objective value < 1 .

C.3.2 Results

Results of the network selectivity analysis are shown in Table C.3. The aldolase (ALD) reaction, the glyceraldehyde 3 phosphate dehydrogenase (GAPDH), and the trypanosomal hexose transporter which is expressed by the blood-stream form seem to have the largest selectivity and should therefore be prioritised as targets in the search for trypanocidal drugs. Experimental

C.3. GLYCOLYSIS IN HUMAN ERYTHROCYTES

Supplementary Table C.2: Tolerated inhibitor concentrations which reduce the flux through glycolysis by 5%.

| Target | MOA | Inhibitor concentration |
|---------|-----|-------------------------|
| BPGP | n | 0.276 |
| BPGP | u | 0.292 |
| HEX | u | 0.640 |
| BPGM | a | 0.840 |
| HK | n | 1.01 |
| HK | k | 1.26 |
| PK | n | 3.03 |
| PFK | c | 3.44 |
| PFK | n | 4.67 |
| BPGP | c | 5.04 |
| PFK | k | 5.71 |
| ENO | n | 5.82 |
| PGM | n | 10.7 |
| THT | n | 16.0 |
| ATPase | a | 22.5 |
| THT | u | 24.7 |
| PGK | n | 44.9 |
| HK | a | 47.4 |
| PGK | c | 53.0 |
| THT | c | 63.7 |
| GPI | n | 90.3 |
| HK | c | 102 |
| PGK | k | 146 |
| PFK | u | 152 |
| GAPDH | n | 177 |
| GAPDH | u | 178 |
| LDHNADH | n | 218 |
| BPGM | n | 255 |
| PGK | u | 289 |
| BPGP | a | 401 |
| GAPDH | k | 402 |
| Phiexch | n | 654 |
| PFK | a | 721 |
| GAPDH | c | 734 |
| ALD | u | 4380 |
| AK | n | 20700 |

results have already shown that decreasing the concentration of ALD and GAPDH is trypanocidal [Cáceres et al., 2010], while for the GAPDH inhibitors selectively targeting the trypanosomal homolog of the enzyme have already been designed [Aronov et al., 1999]. Apart from targets with a high

C.3. GLYCOLYSIS IN HUMAN ERYTHROCYTES

Supplementary Table C.3: Inhibitor concentrations necessary for an effective treatment of *Trypanosoma brucei* and in parallel tolerated by human erythrocytes. The quotient of both concentrations gives the network selectivity defined in Eq. 5.2.

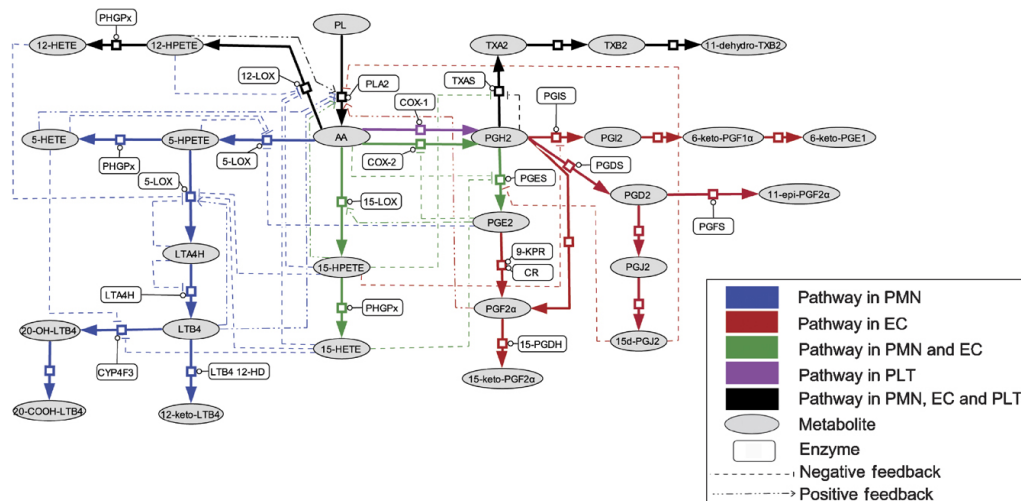
| Target | MOA | Necessary concentrations | | | Network selectivity | |
|---------|-----|--------------------------|--------|------------|---------------------|----------|
| | | Albert | Achcar | Holzhütter | Albert | Achcar |
| HK | k | 1970 | 1980 | 1.26 | 0.000639 | 0.000636 |
| PFK | k | | 946 | 5.71 | | 0.00603 |
| HK | u | 41.7 | 42.2 | 0.64 | 0.0154 | 0.0152 |
| PGK | c | 2780 | 3420 | 53 | 0.019 | 0.0155 |
| PFK | c | 1610 | 177 | 3.44 | | 0.0195 |
| HK | n | 37.5 | 37.9 | 1.01 | 0.027 | 0.0267 |
| ALD | c | | 9150 | 2070 | | 0.226 |
| HK | c | 372 | 378 | 102 | 0.273 | 0.269 |
| PK | n | 9.78 | 9.84 | 3.03 | 0.31 | 0.308 |
| PFK | n | 190 | 14.8 | 4.67 | | 0.315 |
| PGK | k | 273 | 298 | 146 | 0.533 | 0.489 |
| ENO | n | 5.86 | 5.95 | 5.82 | 0.994 | 0.979 |
| PGK | n | 24.9 | 25.1 | 44.9 | 1.8 | 1.79 |
| PGI | n | 27.7 | 28.1 | 90.3 | 3.26 | 3.21 |
| PGM | n | 1.58 | 1.59 | 10.7 | 6.77 | 6.72 |
| PFK | u | 138 | 18.4 | 152 | 1.1 | 8.25 |
| THT | c | 6.46 | 6.53 | 63.7 | 9.85 | 9.75 |
| PGK | u | 25.1 | 25.3 | 289 | 11.5 | 11.4 |
| GAPDH | k | 28 | 31.2 | 402 | 14.4 | 12.9 |
| THT | n | 1.07 | 1.09 | 16 | 14.9 | 14.7 |
| THT | u | 1.28 | 1.29 | 24.7 | 19.3 | 19.1 |
| ALD | n | 7.89 | 8.14 | 282 | 35.8 | 34.7 |
| GAPDH | u | 4.71 | 4.86 | 178 | 37.7 | 36.6 |
| GAPDH | n | 2.51 | 2.82 | 177 | 70.6 | 62.7 |
| GAPDH | c | 5.48 | 6.67 | 734 | 134 | 110 |
| ALD | u | | 7.19 | 4380 | | 609 |
| PT | n | 1.29 | 1.31 | | | |
| PGM | u | 1.58 | | | | |
| GPDH | n | 3.37 | 3.48 | | | |
| GPDH | u | 3.44 | 3.56 | | | |
| ENO | u | 5.86 | | | | |
| TPI | n | 27.9 | 28.5 | | | |
| PGI | u | 27.9 | | | | |
| TPI | u | 29.9 | | | | |
| GPDH | c | 174 | 172 | | | |
| GPDH | k | 204 | 226 | | | |
| AU | n | 209 | 212 | | | |
| TPI | c | 481 | | | | |
| PGI | c | 2990 | | | | |
| PGM | c | 14700 | | | | |
| ENO | c | 25100 | | | | |
| PK | u | | 9.84 | | | |
| PK | k | | 122 | | | |
| PK | c | | 1280 | | | |
| AU | a | | 150 | | | |
| GT | n | | 5160 | | | |
| 3PGAT | n | | 11600 | | | |
| ATPase | n | | | 0.05 | | |
| BPGP | n | | | 0.276 | | |
| BPGP | u | | | 0.292 | | |
| BPGM | a | | | 0.84 | | |
| BPGP | c | | | 5.04 | | |
| ATPase | a | | | 22.5 | | |
| HK | a | | | 47.4 | | |
| LDHNADH | n | | | 218 | | |
| BPGM | n | | | 255 | | |
| BPGP | a | | | 401 | | |
| Phiexch | n | | | 654 | | |
| PFK | a | | | 721 | | |
| AK | n | | | 20700 | | |

selectivity, there are also targets with an infinite selectivity (i.e. the pyruvate transporter and the triose phosphate isomerase). Numerical simulations suggest that the erythrocyte is insensitive to inhibitions of these reactions.

C.4. ARACHIDONIC ACID PATHWAY IN DIFFERENT HUMAN CELLS

Furthermore, the glycerol 3 phosphate oxidase (including the trypanosomal alternative oxidase) the glycerol 3 phosphate dehydrogenase, and some transporters (Gly-3-P DHAP antiporter, 3-phosphoglycerate transporter, glycosomal glucose transporter) are not present in erythrocytes. Thus, they would also make good targets. Apart from these targets some inhibitors have an infinite selectivity because the specific inhibition cannot be included into the erythrocyte model because their specific inhibition kinetics are not available.

C.4 Arachidonic acid pathway in different human cells



Supplementary Figure C.10: Structure of the arachidonic acid pathway in the used models. The models and this Figure have been taken from [Yang et al., 2008].

As a mathematical description of the arachidonic acid pathway in different human cells, I use the models describing the pathway in endothelial cells (EC), platelets (PLT), and polymorphonuclear leukocytes (PMN) as described in [Yang et al., 2008]. These models have been the result of a semantic model search in the partially curated branch of the BioModels Database starting from a model of the AA pathway in PMNs [Yang et al., 2007], which is available from BioModels' curated branch (number 106).

C.4. ARACHIDONIC ACID PATHWAY IN DIFFERENT HUMAN CELLS

Supplementary Table C.4: Relevant data from the literature needed for the construction of the objective function.

| Entity | Attribute | Value | Reference |
|--------------|------------------------|-----------------------------------------------------------------------------------------|--------------------------|
| Bloodvessel | length | $1cm$ | [Yang et al., 2008] |
| | diameter | $50\mu m$ | [Yang et al., 2008] |
| | surface | $1cm \cdot \pi \cdot 25\mu m$ $= 785398\mu m^2$ | |
| | volume | $1cm \cdot \pi \cdot (25\mu m)^2$ $\approx 0.02mm^3$ | |
| Endothelials | surface | $1000\mu m^2$ | [Jaffe, 1987] |
| | number cells in vessel | 785 | |
| | volume (estimated) | $\frac{4\pi}{3}(25\mu m$ $\cdot 10\mu m \cdot 10\mu m)$ $\approx 10000\mu m^3$ | |
| | total volume | $\approx 7850000\mu m^3$ | |
| Platelets | concentration | $250000 \frac{platelets}{mm^3}$ | [Flindt, 2006] |
| | number cells in vessel | 4908 | |
| | diameter | $2 - 3\mu m$ | [Campbell, 1993] |
| | volume | $\frac{4\pi}{3}(\frac{2-3}{2}\mu m)^3$ $\approx 4 - 14\mu m^3$ $\approx 9\mu m^3$ | |
| | total volume | $\approx 44200\mu m^3$ | |
| Neutrophils | concentration | $5000 \frac{1}{mm^3}$ | [Alberts et al., 2007] |
| | number cells in vessel | 98 | |
| | diameter | $12 - 15\mu m$ | [Fujibuchi et al., 2007] |
| | volume | $900 - 1800\mu m^3$ $\approx 1350\mu m^3$ | |
| Eosinophils | concentration | $200 \frac{1}{mm^3}$ | [Alberts et al., 2007] |
| | number cells in vessel | 4 | |
| | diameter | $12 - 17\mu m$ | [Young et al., 2006] |
| | volume | $900 - 2600\mu m^3$ $\approx 1750\mu m^3$ | |
| Basophils | concentration | $40 \frac{1}{mm^3}$ | [Alberts et al., 2007] |
| | number cells in vessel | 1 | |
| | diameter | $10 - 14\mu m$ | [Fujibuchi et al., 2007] |
| | volume | $520 - 1400\mu m^3$ $\approx 960\mu m^3$ | |
| Granulocytes | total volume | $\approx 140000\mu m^3$ | |

C.4. ARACHIDONIC ACID PATHWAY IN DIFFERENT HUMAN CELLS

C.4.1 Setting up the objective function

The three separate models have then been combined using semanticSBML and additional reactions and rules have been added which compute the integral of the total amounts of PGE₂, LTB₄, TXA₂, and PGI₂ using volumes provided in Table C.4. E.g.

$$\begin{aligned} \text{PGE}_{\text{total}}^{\dot{}} &= 7850000 \cdot \text{PGE}_{\text{ec}} + 140000 \cdot \text{PGE}_{\text{pmn}} \\ \text{PGE}_{\text{total}}(t=0) &= 0. \end{aligned}$$

The models describe the dynamics of the arachidonic acid pathway after a stimulus has been given. Yang *et al.* defined the objective that should be fulfilled after treatment as the reduction of the cumulative production of PGE₂ and LTB₄ by 90% after an hour. Furthermore, they included the potential side effect of an imbalance between TXA₂ and PGI₂ with the additional objective of keeping the ratio within 20% of its original value.

In the following I will use a similar objective function, which uses the total production of the corresponding substances as they probably better reflect the changes in variables further downstream of the eicosanoid receptors. The only variable for which this makes a difference is TXA₂, whose dynamical response happens in the first 5 minutes after induction and is therefore not visible in the transient concentration after one hour. The complete objective function then reads

$$\mathcal{X}^2 = \left(\frac{\text{PGE}_{\text{total}}}{567000} \right)^2 + \left(\frac{\text{LTB}_{\text{total}}}{4380000} \right)^2 + \left(\frac{\frac{\text{PGI}_{\text{total}}}{\text{TXA}_{\text{total}}} - 0.00854}{0.00854 \cdot 0.2} \right)^2. \quad (\text{C.6})$$

Unfortunately, the model of Yang *et al.* is highly underdetermined in terms of the amount of experimental data it is fitted to. In order to be able to identify effects of the parameter uncertainty, the authors have published four further parametrisations of the model, which are in agreement with the experimental data. As the different parametrisations partially change the behaviour of the system, individual objective functions have to be constructed for them:

$$\mathcal{X}_2^2 = \left(\frac{\text{PGE}_{\text{total}}}{56083646} \right)^2 + \left(\frac{\text{LTB}_{\text{total}}}{2577000} \right)^2 + \left(\frac{\frac{\text{PGI}_{\text{total}}}{\text{TXA}_{\text{total}}} - 0.813}{0.813 \cdot 0.2} \right)^2 \quad (\text{C.7})$$

$$\mathcal{X}_{3/5}^2 = \left(\frac{\text{PGE}_{\text{total}}}{2744580} \right)^2 + \left(\frac{\text{LTB}_{\text{total}}}{7419152} \right)^2 + \left(\frac{\frac{\text{PGI}_{\text{total}}}{\text{TXA}_{\text{total}}} - 0.0118}{0.0118 \cdot 0.2} \right)^2 \quad (\text{C.8})$$

$$\mathcal{X}_4^2 = \left(\frac{\text{PGE}_{\text{total}}}{38402516} \right)^2 + \left(\frac{\text{LTB}_{\text{total}}}{4566180} \right)^2 + \left(\frac{\frac{\text{PGI}_{\text{total}}}{\text{TXA}_{\text{total}}} - 5.19}{5.19 \cdot 0.2} \right)^2. \quad (\text{C.9})$$

C.4. ARACHIDONIC ACID PATHWAY IN DIFFERENT HUMAN CELLS

C.4.2 Results

Supplementary Table C.5: Targets to reduce LTB_4 production ordered by the respective effective inhibitor concentrations needed to reduce LTB_4 to 10%.

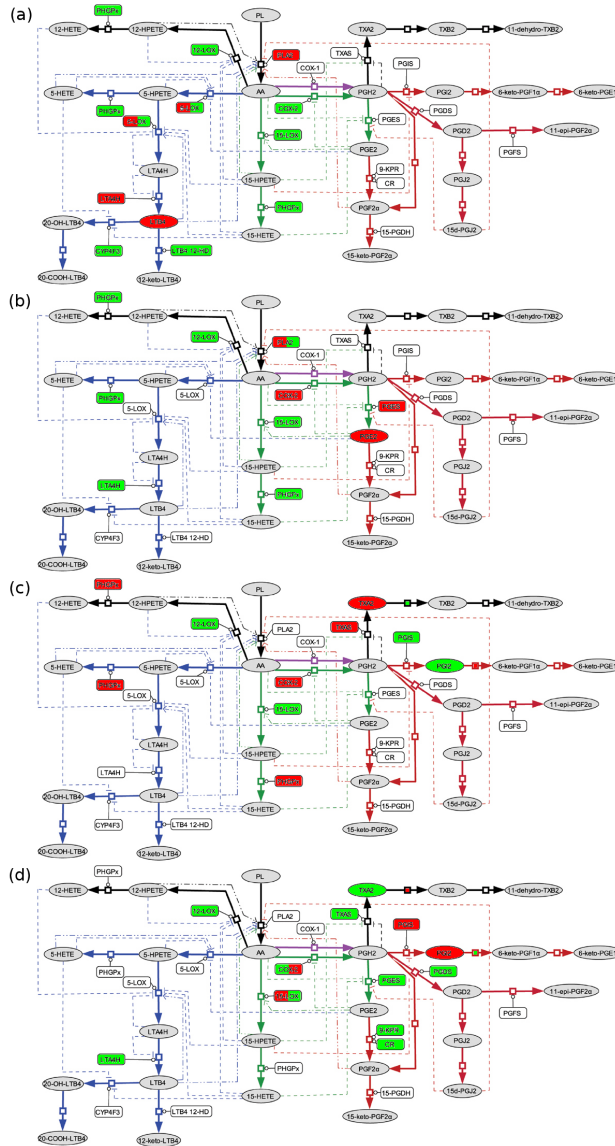
| Type | Target | Reaction | Concentration | | | | |
|------|----------------|-------------|---------------|--------|-------|--------|-------|
| | | | Set 1 | Set 2 | Set 3 | Set 4 | Set 5 |
| a | CYP4F3 | | 3.85 | 3.79 | 2.84 | 4.56 | 2.84 |
| a | LTA4H | degradation | 4.25 | 3.94 | 6.97 | 5.34 | 6.97 |
| n | LTA4H | | 4.93 | 3.22 | 6.82 | 4.96 | 6.82 |
| c | LTA4H | | 10.2 | 6.30 | 7.19 | 6.55 | 7.19 |
| u | LTA4H | | 10.6 | 7.46 | 284 | 21.9 | 284 |
| n | PLA2 | | 14.9 | 37.2 | 4.64 | 50.8 | 4.64 |
| c | PLA2 | | 15.0 | 37.4 | 4.64 | 51.2 | 4.64 |
| a | 15-LOX | | 31.0 | 129 | 8.65 | 7.68 | 8.65 |
| a | 15-LOX | expression | 33.4 | 1830 | 9.19 | 17.0 | 9.19 |
| a | 5-LOX | expression | 52.4 | 251 | 7.18 | 17.6 | 7.18 |
| n | 5-LOX | | 260 | 264 | 4.39 | 12.7 | 4.39 |
| c | 5-LOX | | 292 | 275 | 4.51 | 13.6 | 4.51 |
| a | LTB_4 | degradation | 2500 | 274 | 1560 | | 1560 |
| a | COX-2 | | 5978 | 7170 | 17900 | 349 | 17900 |
| u | PLA2 | | 8860 | 7940 | 5480 | 7830 | 5480 |
| u | 5-LOX | | 15300 | 53500 | 344 | 426 | 344 |
| a | PHGPx | | 34300 | 64600 | 67.1 | 94.4 | 67.1 |
| a | LTA4 | degradation | 66000 | 6690 | 61400 | 310000 | 61400 |
| a | 12-LOX | | 121000 | 228000 | | | |
| a | AA | degradation | 371000 | | 25500 | 6800 | 25500 |
| a | LTB_4 | degradation | | | | 27400 | |

C.4. ARACHIDONIC ACID PATHWAY IN DIFFERENT HUMAN CELLS

Supplementary Table C.6: Targets to reduce PGE₂ production ordered by the respective effective inhibitor concentrations needed to reduce PGE₂ to 10%.

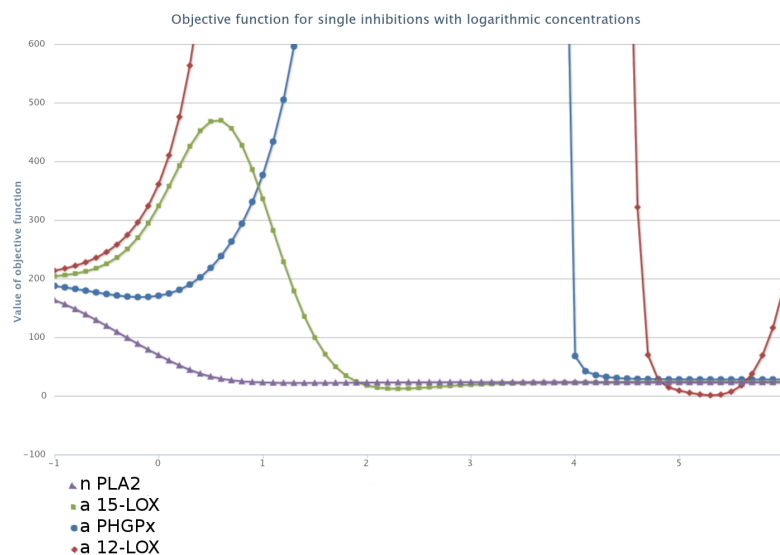
| Type | Target | Reaction | Concentration | | | | |
|------|--------|-------------|---------------|-------|-------|-------|-------|
| | | | Set 1 | Set 2 | Set 3 | Set 4 | Set 5 |
| a | LTA4H | | 7.92 | 6.25 | | 19.9 | |
| n | PLA2 | | 14.8 | 4.36 | 9.80 | 40.1 | 9.80 |
| c | PLA2 | | | 4.68 | | | |
| n | LTA4H | degradation | | 9.94 | | | |
| c | LTA4H | degradation | | 12.5 | | | |
| n | PGES | | 50.7 | 383 | 1230 | 11.2 | 1230 |
| c | PGES | | 50.7 | 413 | 1230 | 12.9 | 1230 |
| a | LTA4 | degradation | | 406 | | | |
| n | COX-2 | | 426 | 25.5 | 79.3 | 77.2 | 79.3 |
| c | COX-2 | | 547 | 25.8 | 90.7 | 114 | 90.7 |
| a | 15-LOX | | 2160 | 31.5 | 233 | 73.1 | 233 |
| a | 15-LOX | expression | 2170 | 60.8 | 235 | 174 | 235 |
| u | PLA2 | | 4660 | 935 | 11400 | 6110 | 11400 |
| u | COX-2 | | 8010 | 3450 | 1240 | 551 | 1240 |
| a | 12-LOX | | 9940 | 21500 | | | |
| a | PLA2 | | 167000 | 84600 | 14200 | 928 | 14200 |
| a | PHGPx | | | 748 | | 19.1 | |
| n | TXAS | degradation | | 1460 | | | |
| u | LTA4H | degradation | | 2070 | | | |
| n | 5-LOX | expression | | 3230 | | | |
| u | PGES | | | 7030 | | 195 | |
| a | AA | degradation | | 9000 | | | |
| a | TXAS | | | | | 13900 | |

C.4. ARACHIDONIC ACID PATHWAY IN DIFFERENT HUMAN CELLS



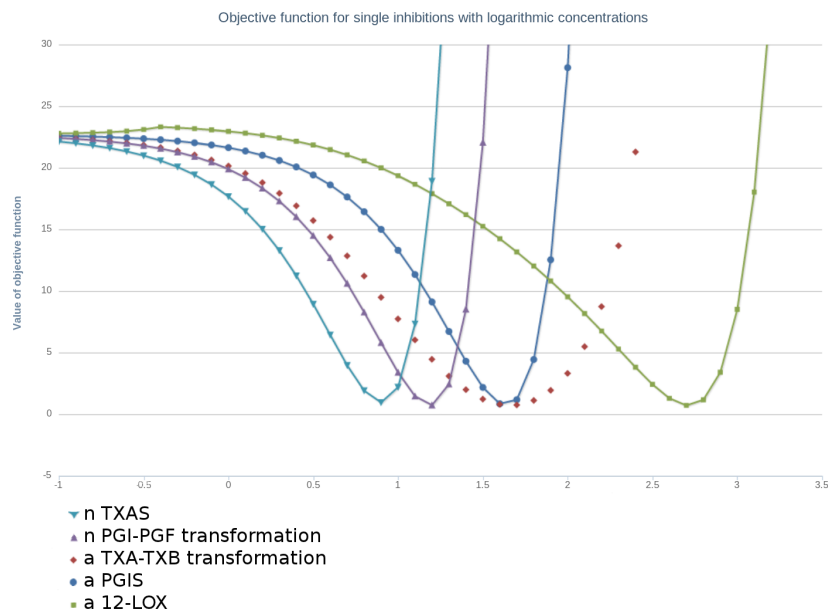
Supplementary Figure C.11: Targets with effects on various variables for the first parameter set. Red and green ellipses signify the compounds the should be reduced or increased in their concentration. Red and green rectangles show which inhibitions or activations lead to the desired effect, where the inhibitors and activators are also allowed to affect the enzyme levels by changing its expression or degradation. The figures show from top to bottom the targets being able to lower (a) LTB₄ or (b) PGE₂ production to 10% or to (c) increase or (d) decrease the PGI₂/TXA₂ ratio by a factor of 2. The network structure image is again taken from [Yang et al., 2008].

C.4. ARACHIDONIC ACID PATHWAY IN DIFFERENT HUMAN CELLS



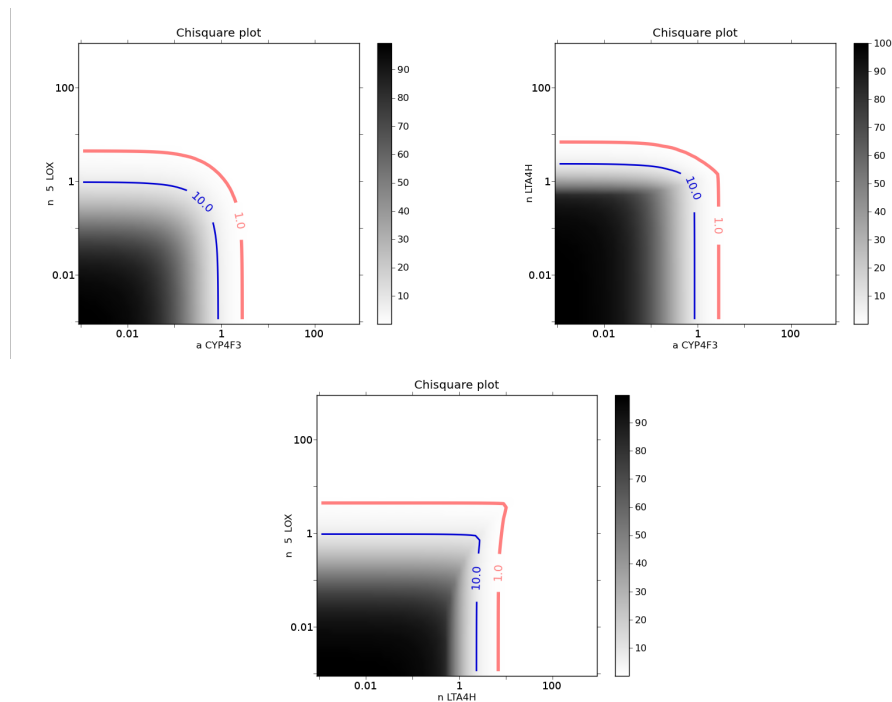
Supplementary Figure C.12: Effects of varying inhibitor concentrations on the objective function. Results show a noncompetitive inhibitor of the phospholipase A2 and non-essential activators of 15- and 12-lipoxygenase and phospholipid hydroperoxide glutathione peroxidase. These inhibitors can potentially reduce LTB_4 and PGE_2 levels in parallel and could therefore be potential single drugs.

C.4. ARACHIDONIC ACID PATHWAY IN DIFFERENT HUMAN CELLS



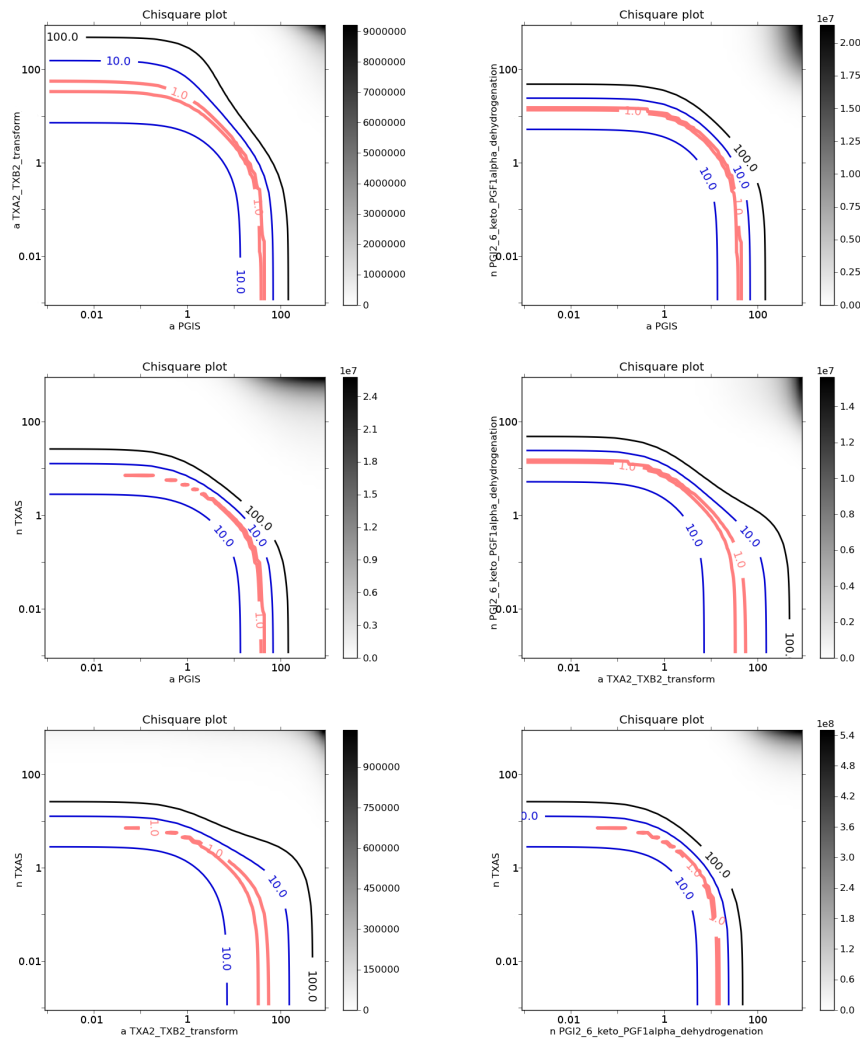
Supplementary Figure C.13: Effects of different effective inhibitor concentrations on the objective function when the system is in parallel treated using a non-competitive inhibitor for phospholipase A2 with an effective concentration of 100.

C.4. ARACHIDONIC ACID PATHWAY IN DIFFERENT HUMAN CELLS



Supplementary Figure C.14: Nonidentifiabilities between targets reducing LTB_4 production. The plots show the objective function in dependence to varying concentrations of non-competitive inhibitors of 5-LOX and LTA4H and a non-essential activator of CYP4F3. In parallel the system is treated with a non-competitive inhibitor of PGES with an effective concentration of 10000. The plots show that for every concentration of one of the non-identifiable drugs there is a range of concentrations of a different inhibitor compensating for it. In contrast to other examples this simulations are performed using the third parameter set as the non-competitive inhibitor of 5-LOX does not work for the first parametrisation.

C.4. ARACHIDONIC ACID PATHWAY IN DIFFERENT HUMAN CELLS



Supplementary Figure C.15: Nonidentifiabilities between targets increasing PGI_2 over TXA_2 levels. Plots again show the objective function in dependence to varying inhibitor and activator concentrations. Apart from the drugs shown on the axes the system is treated with a non-competitive inhibitor of PLA_2 with an effective concentration of 100. The Figure signifies that PGI_2 over TXA_2 level increasers are non-identifiable when applied in conjunction with an inhibitor of PLA_2 .

C.4. ARACHIDONIC ACID PATHWAY IN DIFFERENT HUMAN CELLS

Supplementary Table C.7: List of solutions provided by the Yang *et al.* using the MTOI method. Lines represent clusters of possible solutions, where the letter y signifies that a drug against this target is used in the solution, the frequency denotes in how many parametrisations this solution has been observed, and S denotes the sensitivity of the objective function with respect to the inhibitor concentrations. The remaining two columns describe the final solutions provided by Yang *et al.* and the results in chapter 5. A “b” denotes here that this is a basis solution provided by the respective author and “d” denotes that this solution can be derived from basis solutions as it is a superset of them. The basis solutions of Yang *et al.* cover 14 of 23 found solutions, while the basis solutions identified in chapter 5 cover 20 of them. These basis solutions do not only represent the all found solutions better but are, in contrast to the solutions provided by Yang *et al.* , mathematically motivated and reproducible.

| Targets | | | | | | | Solutions | |
|---------|-----|------|-------|-------|-----------|--------|-----------|------|
| PLA2 | COX | PGES | 5-LOX | LTA4H | Frequency | S | MTOI | TIde |
| y | y | - | - | - | 1 | 0.0011 | - | b |
| y | y | y | - | - | 1 | 0.0011 | - | d |
| - | y | y | y | y | 5 | 0.002 | d | d |
| - | y | y | - | y | 5 | 0.0021 | d | d |
| - | - | y | y | y | 2 | 0.0026 | d | d |
| y | y | y | y | y | 5 | 0.0028 | d | d |
| - | y | y | y | - | 5 | 0.0033 | - | d |
| y | y | y | - | y | 5 | 0.0034 | d | d |
| - | y | - | y | y | 5 | 0.0036 | d | d |
| y | y | - | y | y | 5 | 0.0037 | d | d |
| y | y | y | y | - | 5 | 0.0041 | d | d |
| y | y | - | - | y | 5 | 0.0042 | b | d |
| - | y | - | - | y | 5 | 0.0048 | b | b |
| y | - | y | y | y | 1 | 0.0057 | d | d |
| - | - | y | y | - | 2 | 0.0061 | - | b |
| - | - | y | - | y | 3 | 0.0072 | b | b |
| y | y | - | y | - | 5 | 0.0075 | b | d |
| - | y | - | y | - | 5 | 0.01 | - | b |
| y | - | y | - | y | 1 | 0.0123 | d | d |
| y | - | y | y | - | 1 | 0.0241 | - | d |
| y | - | - | y | y | 1 | 0.0431 | - | - |
| y | - | - | - | y | 1 | 0.0569 | - | - |
| y | - | - | y | - | 1 | 0.0624 | - | - |

Abstract

Over the last decade the productivity of the pharma industry has been constantly declining. Less and less drugs against new diseases are admitted to the market each year. This is mainly due to the fact that an increasing number of drug candidates fail for a lack of *in vivo* activity or for their toxicity in clinical trials.

In order to reduce this failure rate, the targets against which new drugs are developed have to be chosen more carefully. This can be done with the help of methods from Systems Biology with which the dynamical effects of hypothetical drugs can be modelled *in silico*. The combination of mathematical models with experimental data will improve the target selection and will make the resulting drugs less likely to fail in clinical trials.

Within this work I have developed a framework for the application of kinetic models in the drug development process. Furthermore, I have developed methods and tools that support researchers in pursuing the framework. This includes methods for the automated retrieval of mathematical models that describe processes relevant to an investigated disease, methods for the integration of knowledge stored in these models, and the investigation of the combined information for potential drug targets.

For the prioritisation of drug targets I propose different objectives and methods. Depending on the diseases, one can either choose to only consider the efficacy of drugs against potential targets or one can decide to incorporate information on potential side-effects in the considered or in alternative models. These objective can then be used in exhaustive searches for optimal combinations of hypothetical drugs. Apart from this identification of optimal treatments, I introduce methods that allow for the discovery of treatment alternatives, which can be useful when drugs against a selected target are hard or even impossible to create. Furthermore, I discuss methods for the investigation of synergisms and antagonisms amongst hypothetical drugs. Knowledge about these drug combination effects can be exploited to create treatments with fewer side-effects or treatments against which resistances are less likely to develop.

In order to prove the relevance of the investigated methods, these are applied to two example systems, the glycolysis in *Trypanosoma brucei*, the pathogen causing the African sleeping sickness, and the arachidonic acid pathway in different human cells. The obtained results generally agree with the knowledge available in the literature but extend the understanding of drug effects on these networks.

Zusammenfassung

Über die letzten Jahre ist die Produktivität der Pharmaindustrie deutlich zurück gegangen. Jedes Jahr werden weniger Medikamente zum Markt zugelassen und dies liegt hauptsächlich daran, dass viele Kandidaten aufgrund von zu geringer Effizienz oder wegen ihrer Toxizität durch klinische Studien fallen.

Um diese hohe Ausfallrate zu verringern, sollten die Medikamententargets, die von den Kandidaten angegriffenen Proteine, sorgsamer ausgewählt werden. Ein Weg, dies zu bewerkstelligen, führt über die mathematische Modellierung von biologischen Systemen. Über derartige Modelle von krankheitsrelevanten Stoffwechsel-, Signaltransduktions- oder Genregulationsnetzwerken ist es möglich den Effekt eines Medikamentes in Computersimulationen vorherzusagen und so zu wirksamen und sicheren Targets zu gelangen.

In dieser Arbeit habe ich ein Framework für die Anwendung von mathematischen Modellen in Form von Differentialgleichungen für die Medikamentenforschung entwickelt. Des Weiteren habe ich Methoden und Software entwickelt, die Forschern bei der Entwicklung und Auswertung von mathematischen Modellen in diesem Zusammenhang unterstützen. Dies umfasst Methoden zur Suche nach existierenden Modellen von krankheitsrelevanten Pathways, deren Integration und die Simulation von hypothetischen Medikamenten in ihnen.

Für das Auffinden von optimalen Targets in mathematischen Modellen habe ich unterschiedliche Kriterien und Methoden entwickelt. Innerhalb meines Frameworks lassen sich hypothetische Medikamente sowohl nach ihrer Effizienz bewerten, als auch nach möglichen Nebeneffekten im betrachteten oder anderen mathematischen Modellen. Für die Identifikation von Zielen, die nach diesen Kriterien optimal sind, habe ich verschiedene Methoden entwickelt. Während die erste Methode eine erschöpfende Suche über mögliche Medikamentenkombinationen durchführt, identifiziert die zweite Methode Targets, die zu gleichen Effekten führen und daher Behandlungsalternativen darstellen. Die dritte Methode untersucht Medikamentenkombinationen auf synergistische oder antagonistische Kombinationswirkungen, da diese zu

nebenwirkungsarmen Behandlungen oder Behandlungen, gegen die die Resistenzbildung verlangsamt, wird führen können.

Um die Anwendbarkeit meiner Methoden zu demonstrieren, wende ich sie auf zwei Beispielsysteme an, die Glykolyse im Pathogen *Trypanosoma brucei*, dem Erreger der Schlafkrankheit, und den Arachidonsäure Pathway in menschlichen Zellen. Die erhaltenen Ergebnisse decken sich größtenteils mit bekanntem Wissen, sie erlauben jedoch einen detaillierteren Einblick in die Wirkungsweise von erfolgreichen Behandlungen.

Curriculum vitae

Der Lebenslauf ist in der Online-Version aus Gründen des Datenschutzes nicht enthalten.

Selbstständigkeitserklärung

Hiermit erkläre ich, dass ich diese Arbeit selbstständig verfasst habe und keine anderen als die angegebenen Quellen und Hilfsmittel in Anspruch genommen habe. Ich versichere, dass diese Arbeit in dieser oder anderer Form keiner anderen Prüfungsbehörde vorgelegt wurde.

Berlin, September 2012