Integration of Ligand Characteristics for the Simulation of Cellular Reactions

Dissertation zur Erlangung des akademischen Grades
doctor rerum naturalium (Dr. rer. nat.)
im Fach Bioinformatik

eingereicht im
Fachbereich Mathematik und Informatik
der Freien Universität Berlin

vorgelegt von
Diplom-Informatiker Raphael André Bauer

Mai 2009

Tag der Disputation: 10. Dezember 2009

Betreuer:	Priv.-Doz. Dr. Robert Preissner Charité Universitätsmedizin Berlin
Betreuer am Fachbereich:	Prof. Dr. Martin Vingron Freie Universität Berlin Max-Planck Institut für molekulare Genetik
Externer Gutachter:	Prof. Dr. Ralf Zimmer Ludwig-Maximilians-Universität München

Abstract

A major characteristic of life sciences is the generation of vast amounts of raw data produced by modern wet lab technologies. The data may come from large-scale experiments where chemical compounds are tested on their ability to act as possible agents against certain diseases. It can also originate in the determination of the 3D structure of macromolecules, or from the genetic code of a species. Evaluation and integration of that raw data is becoming increasingly important. Currently, often only a fraction of the generated data is integrated and evaluated.

The data integration problem is addressed in the first part of the work. A data-warehouse is developed that integrates 3D information on proteins with information about potential drugs, potential binding sites and advanced 3D binding site screening techniques. Furthermore, as similarity screening of molecules and proteins can often only be carried out with limited accuracy on a limited amount of data sets. A framework is presented that facilitates the integration of data sources and methods with an emphasis on exact 3D screening techniques.

The amount of searchable macromolecular structures, such as proteins and RNAs is growing rapidly. However, there are only a few methods available allowing for a rapid 3D screening of thousands of proteins, and only a handful of methods can be used for aligning RNA structures. A novel method is presented that uses n-grams and index structures in concert with a nomenclature that reduces a biomolecule to a string. It can be shown that the method delivers comparable or better results in comparison to leading methods in the field of protein and RNA alignment. The method can be used in high throughput experiments because of its precision and adjustable speed.

The last part of the work deals with interaction and signal transduction. Expression levels correlate to the amount of signals that are transduced in a biological network. Various concepts are evaluated that map expression levels of genes on the apoptosis signal transduction network using Petri nets. Finally, a software package is presented that is able to simulate Petri nets based on the developed paradigm. The software can hide the complexity of the Petri net, which allows non-computer experts to use the software efficiently.

Keywords: Drug similarity, substructure search, structural alignment, protein, RNA, hash-table, n-grams, dihedral angles, Petri net, microarray, apoptosis

Science must become art.

Carl von Clausewitz

Publications

The majority of this thesis has been published in peer reviewed journals. Please find the articles corresponding to the parts of this thesis below.

- Bauer R. A., Günther S., Heeger C., Jansen D., Thaben P. and Preissner R. (2008). SuperSite: Dictionary of metabolite and drug binding sites in proteins. *Nucleic Acids Research*, 37 (Database issue): D195-D200. [Section 2.1]

- Bauer R. A., Bourne P. E., Formella A., Frömmel C., Gille C., Goede A., Guerler A., Hoppe A., Knapp E. W., Poschel T., Wittig B., Ziegler V. and Preissner R. (2008). Superimpose: a 3D structural superposition server. *Nucleic Acids Research*, 36 (Web Server issue): W47-W54. [Section 2.2]

- Bauer R. A., Rother K., Bujncki J. M. and Preissner R. (2008). Suffix techniques as a rapid method for RNA substructure search. *Genome Informatics*, 20: 183-198. [Section 3.1]

- Bauer R. A., Rother K., Moor P., Reinert K., Steinke T., Bujnicki J. M., Preissner R. (2009). Fast structural alignment of biomolecules using a hash table, n-grams and string descriptors. *Algorithms*, 2(2), 692-709. [Section 3.2]

- Hildebrand P., Goede A., Gruening B., Michalsky E., Bauer R. A., Ismer J., Preissner R. (2009). SuperLooper - A prediction server for the modeling of loops in globular and membrane proteins, *Nucleic Acids Research*, 37 (Web Server issue), (accepted). [Section 1.2.2]

Software

Most of the software packages developed in the course of this thesis are freely available as open source.

- lrrr, http://lrrr.sf.net [Section 2.1]
- suiteRNA, http://suiterna.sf.net [Section 3.1]
- LaJolla, http://lajolla.sf.net [Section 3.2 and Section 3.3]

Online resources

The online resources are hosted at the Charité Computing Center, CBF. The Structural Bioinformatics Group of the Charité updates and maintains the data sources.

- SuperSite, http://bioinformatics.charite.de/supersite [Section 2.1]
- Superimposé, http://bioinformatics.charite.de/superimpose [Section 2.2]
- SuperRNAAlign, http://bioinformatics.charite.de/superrnaalign [Section 3.3]

Acknowledgements

The projects leading to this thesis were carried out at the Structural Bioinformatics Group of the Department for Molecular Biology and Bioinformatics / Physiology of the Charité Medical University, Berlin.

First and foremost, I thank Robert Preissner for giving me the opportunity to work at the Structural Bioinformatics Group and to study at the graduate school, as well as for supervising this thesis. His kind patience, excellent support and guidance along the way were invaluable for the success of this project. Special thanks go to Martin Vingron for monitoring and supportively advising my projects as member of my PhD committee and graduate school and especially to Ralf Zimmer for refereeing this PhD thesis.

Without the scientific expertise and cooperation of Kristian Rother and Stefan Günther as well as Dorothee Biener and Max Liebl this work would not have been possible.

I was privileged to be accepted as student of the international graduate school "Genomics and Systems Biology of Molecular Networks – Berlin, Boston, Kyoto". I am really grateful for the support of the Principal Investigators who organized and attended conferences, gave critical feedback and opened my eyes to the stunning perspectives of our era: Oliver Ebenhöh, Martin Falcke, Hanspeter Herzel, Hermann-Georg Holzhütter, Edda Klipp, Ernst-Walter Knapp and Nikolaus Rajewsky. I especially want to thank my fellow students of the graduate school for making this PhD journey such a great experience both in terms of excellent scientific exchange as well as friendship.

Chapter 2 was additionally supported by Philip E. Bourne, Arno Formella, Cornelius Frömmel, Christoph Gille, Andrean Goede, Aysam Guerler, Carolin Heeger, Andreas Hoppe, Dominic Jansen, Thorsten Pöschel, Paul Florian Thaben, Burghardt Wittig and Valentin Ziegler. Chapter 3 was supported by Peter Moor, Knut Reinert, Thomas Steinke and Janusz M. Bujnicki. Chapter 4 would not have been possible without the enthusiasm of Konstantin Pentchev and Christian Scholz who successfully completed their bachelor and master's theses in the area. Invaluable advice regarding Petri net theory came from Ina Koch. The net architecture would not have been possible in such great detail

without the excellent expertise of Jessica Ahmed, Julia Hossbach and Peter Daniel. Additional support came from Robert Adams and Robert Lehmann.

The special guidance of Felix Bübl, Peter Hildebrand and Ulf Leser was invaluable.

I thank all former and current members of the Structural Bioinformatics Group for the excellent working atmosphere, for scientific discussions and for many enjoyable after-work activities. Apart from the members already mentioned I would like to thank: Mathias Dunkel, Joachim von Eichborn, Melanie Füllbeck, Ines Jäger, Elke Michalsky, Ulrike Schmidt and Swantje Struck. I also thank the students with whom I had the opportunity to work: Björn Grüning, Nico Güttler and Marcus Schroeder. Very special thanks go to Claude Couture, Franz Brendl, Peter Dulovits, Sebastian Heglmeier, Thomas Meister, Rebecca Miller and Christian Suchy.

My deepest thanks go to my family, especially my parents Harry and Ilse, my brother Severin and Cathrin for their true love, care and support.

Without them and the many people that I have not mentioned here by name, this work would never have been possible.

Contents

Chapter 1

Introduction

1.1 A greater picture of systems biology

1.1.1 A declaration and a virtual human

“Recent advances in systems biology indicate that the time is now ripe to initiate a grand and challenging project to create a comprehensive, molecules-based, multi-scale, computational model of the human (the virtual human) body over the next thirty years, capable of simulating and predicting, with a reasonable degree of accuracy, the consequences of most of the perturbations that are relevant to health-care.” (The Tokyo Declaration, Future Challenges for Systems Biology, Tokyo International Forum, Tokyo, February 4–6, 2008).

The Tokyo Declaration briefly summarized what systems biology should be about: Building a complete model of a complex life form such as a human being [Kitano, 2001,Kitano, 2002,Kell, 2007]. The objective is to understand how things inside us really work. And even more important: To understand the source of diseases and to find cures with minimal side effects and maximum effect [Scheiber et al., 2009]. According to Hiroaki Kitano, the vision of the virtual human is to be realized by around 2050 [News, 2008]. This thesis attempts to bring together some small pieces of the puzzle as one of the many steps required to make this vision happen.

1.1.2 Double helix – single sequence

The father of modern biology was Charles Darwin, who did not accept that everything is created by a higher being (Figure 1.1). Darwin proposed that life forms evolve passing on phenotypes and attributes to their descendants [Bowler, 2009]. He came to his conclusions without exact knowledge of DNA (deoxyribonucleic acid) or molecular biology [Berkman et al., 2008]. In the last

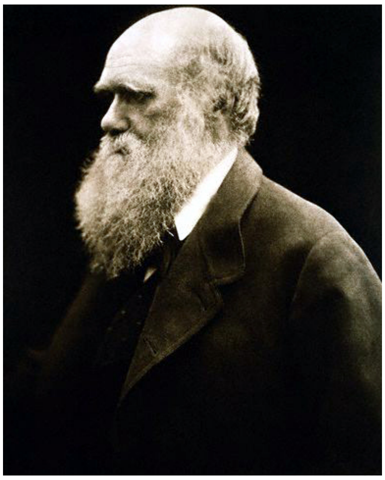


Figure 1.1: A portrait of Charles Darwin by Julia Margaret Cameron. (Picture source: Wikipedia, copyright expired)

century, this gap was filled with the discovery of the double helix, the principle of semiconservative replication and the knowledge of the properties of nucleotides [Watson and Crick, 1953]. The ability to sequence these nucleotides with methods based on Sanger's discovery [Sanger et al., 1977] ultimately led to the publication of the first consensus sequence of the human genome [Morgan, 2001, Venter et al., 2001]. This consensus sequence was only possible due to advances in biotechnical sciences, as well as faster computational capabilities. Now the focus of research has shifted again. The sequence of the human genome has been deciphered, but not decoded, and many questions are left open. Many questions concern the structure of proteins built from information in DNA. Other open questions concern the regulatory mechanism of a life form that is influenced by the interplay of proteins and DNA. During the last 50 years numerous breakthroughs in biotechnical sciences were made toward determining the 3D structure of a biomolecule, such as RNA (ribonucleic acid) and proteins. And still there is a huge gap between the number of proteins known to exist and the number of 3D structures of proteins that are available. For many important proteins, it is still not possible to determine their 3D structures due to the limitations of the methods [Park et al., 2008]. More interestingly, sequencing and determining 3D structures are generating terabytes of data that have to be connected and evaluated. The evaluation can be e.g. on sequential level or based on 3D structures. The core disciplines dealing with the data are bioinformatics and systems biology. They are using the power of computers to generate models and predictions based on a huge heterogeneous amount of information. These predictions can be on small scale interactions of elements of a cell but also on a bigger scale like interactions between cells or organs. The final goal is to have all information and interactions interconnected, leading to the virtual human outlined in the Tokyo Declaration.

1.1.3 From sequence to regulation

In the late nineties of the twentieth century, expectations from the complete consensus sequence of a human genome were huge (Figure 1.2). It was commonly thought that knowledge of the sequence would automatically lead to the complete understanding of life. But this view had to be revised [Stein, 2008]. After the initial excitement it became clear that the genome is important, but only explains a fraction of how life works. The finished sequence of a human genome brought up more questions than it could answer.

The process of reading genes involves an intermediate step where DNA is transcribed into RNA that is then translated into proteins. A complex regulatory process is active in order to form different cells originating from one single cell in multi-cellular organisms. Life is all about interacting elements and not only about a one-dimensional DNA sequence. The initial enthusiasm about the release of the consensus sequence of the human genome quickly turned into realism and to novel questions and challenges that have to be addressed. A new discipline that attempts to shed light on the regulatory mechanisms on all levels of the cellular apparatus is systems biology.

1.1.4 Regulation and interaction in the cellular apparatus

Regulation and cellular response can be measured with a range of ever evolving biotechnical methods [Yengi, 2005]. The purpose of these methods is e.g. to figure out if elements of a cell are active in a certain context. Theoretically it is possible to conclude the amount of protein produced from the amount of transcribed RNA of a gene. Measuring RNA is simple and inexpensive and can be done using technologies such as microarrays [Keller et al., 2008]. However, this indirect measurement is often inaccurate because not all RNA is translated into protein. Measuring the amount of protein directly is more precise and is done increasingly nowadays. The direct measurement is still much more time-consuming and expensive [Cohen et al., 2008]. The composition of a cell in terms of RNA and protein differs greatly depending on the tissue of origin [Noble, 2002]. Furthermore, pathologically altered cells such as cancer cells have a different fingerprint in terms of RNA and protein. This fingerprint can be used in diagnostic methods like methylation analysis leading

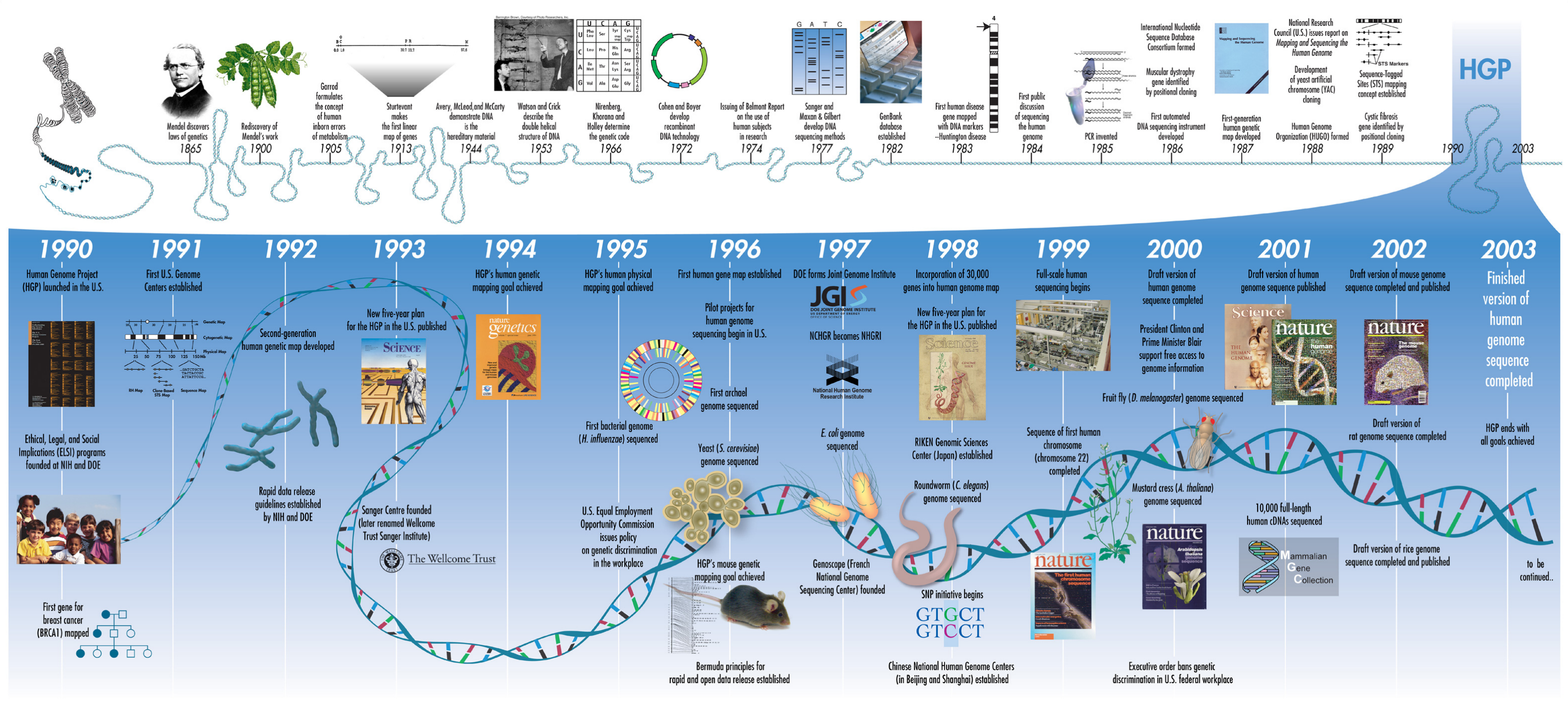

Mendel discovers laws of genetics 1865
Rediscovery of Mendel's work 1900
Garrod formulates the concept of human inborn errors of metabolism 1905
Sturtevant makes the first linear map of genes 1913
Avery, McLeod,and McCarty demonstrate DNA is the hereditary material 1944
Watson and Crick describe the double helical structure of DNA 1953
Nirenberg, Khorana and Holley determine the genetic code 1966
Cohen and Boyer develop recombinant DNA technology 1972
Issuing of Belmont Report on the use of human subjects in research 1974
G A T C
Sanger and Maxan & Gilbert develop DNA sequencing methods 1977
GenBank database established 1982
First human disease gene mapped with DNA markers --Huntington disease 1983
First public discussion of sequencing the human genome 1984
PCR invented 1985
International Nucleotide Sequence Database Consortium formed
Muscular dystrophy gene identified by positional cloning
First automated DNA sequencing instrument developed 1986
First-generation human genetic map developed 1987
National Research Council (U.S.) issues report on *Mapping and Sequencing the Human Genome*
Development of yeast artificial chromosome (YAC) cloning
Human Genome Organization (HUGO) formed 1988
STS Markers
Sequence-Tagged Sites (STS) mapping concept established
Cystic fibrosis gene identified by positional cloning 1989
HGP
1990
2003
1990
Human Genome Project (HGP) launched in the U.S.
Ethical, Legal, and Social Implications (ELSI) programs founded at NIH and DOE
First gene for breast cancer (BRCA1) mapped
1991
First U.S. Genome Centers established
Genetic Map
Cytogenetic Map
Physical Map
RH Map
Clone-Based STS Map
Sequence Map
1992
Second-generation human genetic map developed
Rapid data release guidelines established by NIH and DOE
1993
New five-year plan for the HGP in the U.S. published
Science
Sanger Centre founded (later renamed Wellcome Trust Sanger Institute)
The Wellcome Trust
1994
HGP's human genetic mapping goal achieved
nature genetics
1995
HGP's human physical mapping goal achieved
First bacterial genome (*H. influenzae*) sequenced
U.S. Equal Employment Opportunity Commission issues policy on genetic discrimination in the workplace
1996
First human gene map established
Pilot projects for human genome sequencing begin in U.S.
First archael genome sequenced
Yeast (*S. cerevisiae*) genome sequenced
HGP's mouse genetic mapping goal achieved
Bermuda principles for rapid and open data release established
1997
DOE forms Joint Genome Institute
JGI
DOE JOINT GENOME INSTITUTE
US DEPARTMENT OF ENERGY
OFFICE OF SCIENCE
NCHGR becomes NHGRI
National Human Genome Research Institute
E. coli genome sequenced
Genoscope (French National Genome Sequencing Center) founded
1998
Incorporation of 30,000 genes into human genome map
New five-year plan for the HGP in the U.S. published
Science
RIKEN Genomic Sciences Center (Japan) established
Roundworm (*C. elegans*) genome sequenced
SNP initiative begins
GTGCT
GTCCT
Chinese National Human Genome Centers (in Beijing and Shanghai) established
1999
Full-scale human sequencing begins
Sequence of first human chromosome (chromosome 22) completed
nature
2000
Draft version of human genome sequence completed
President Clinton and Prime Minister Blair support free access to genome information
Fruit fly (*D. melanogaster*) genome sequenced
Mustard cress (*A. thaliana*) genome sequenced
nature
Executive order bans genetic discrimination in U.S. federal workplace
2001
Draft version of human genome sequence published
Science
nature
10,000 full-length human cDNAs sequenced
Mammalian Gene Collection
2002
Draft version of mouse genome sequence completed and published
nature
Draft version of rat genome sequence completed
Draft version of rice genome sequence completed and published
2003
Finished version of human genome sequence completed
HGP ends with all goals achieved
to be continued..

Figure 1.2: Prominent figures like Lincoln Stein predicted that once the human genome is deciphered it will render disciplines like Bioinformatics obsolete. (Picture: DarrylLeja, National Human Genome Research Institute; with permission).

to an early diagnosis of diseases such as cancer [Bonetta, 2008]. The response of a cell towards a perturbation such as drugs (small compounds) is also important. Measuring the response to a certain treatment allows for conclusions if the treatment is specific or unspecific for a certain cell type and disease. An unspecific compound can have negative effects on a wide variety of cells, causing adverse drug effects (ADRs) [Scheiber et al., 2009]. The final goal is a working model of known responses and interactions where perturbations can be predicted, leading to a better understanding of the cellular apparatus. This can be used for the development of new drugs where simulated eliminations of interacting elements allow for evaluation of cellular response [Butcher, 2005, Davidov et al., 2003].

1.1.5 Personalized Medicine

With the advent of cheaper sequencing technology an important shift in sequencing is currently taking place [Mardis, 2008]. The focus is shifting from the investigation of consensus genomes of a species to the genome of an individual or the genome of single pathological cells. In medical sciences the switch to the genome and from this to the metabolism of an individual is commonly referred to as Personalized Medicine. Personalized Medicine potentially allows for a better treatment of individual patients, taking into account their individual susceptibility to side effects of certain drugs based on their predicted metabolism [Scheiber et al., 2009]. It will also allow determination of individual risk factors. This issue is currently debated controversially as commercial enterprises first use this technology [Kaye, 2008]. The final goal is to combine the knowledge of the regulation and its effects on individuals in order to provide better directed treatment.

Systems biology is the discipline that emerged from the growing knowledge about regulatory mechanisms of cells. Although there are various definitions of the term, it becomes clear that the discipline has to deal especially with the following fields in order to make the vision of the virtual human come true:

- The integration of vast amounts of heterogeneous data.
- The individual elements from drugs to macromolecules, as well as their interactions.
- The simulation and evaluation on network level.

1.2 Loose ends

The term *loose ends* originates in the old profession of rope making. An important task for rope makers was to finish a rope by securely binding together the (loose) ends. An imperfectly executed final step potentially leads to the disintegration of the whole rope. This is also true for current research in the field of life sciences where a number of loose ends exist that must be tied up. This section is dedicated to the loose ends this thesis is attempting to bind together.

Jch bin ain ſchalk vnd öch ain luder·
vnd würk zemen flachs vnd kuder·
vnd hān es mīt hanf über zogen·Da
mit hān ich die lüt betrgen

Figure 1.3: The text, written in an old German dialect, states that the rope maker cheated his customers by exchanging the better and more expensive hemp inside a rope with the cheaper flax (Picture source: Wikipedia, copyright expired).

1.2.1 An integration gap: elements, processes and methods

Section 1.1 mentioned that the knowledge about many elements and processes of life forms is growing quickly. However, the knowledge and the data is heterogeneous, making the integration of different data a difficult task. An example is the growing knowledge about macromolecular structures such as proteins. Proteins interact with other proteins, or with small compounds. Many proteins are badly characterized regarding their function, and some are even completely uncharacterized. The determination of the function of a protein today includes the combination of a range of information sources. The information sources can be e.g. interaction partners, similarities to characterized proteins in structure and sequence as well as integration of drug and pathway data. This can lead to proposals about wanted and unwanted effects of drugs and is highly important [Editorial, 2009]. As this determination is often only possible using complicated custom tools, it is not carried out regularly.

1.2.2 Biomolecules, string encoding and applications

Growing importance of RNA structure

Section 1.1.3 covers the issue that RNA is not solely a carrier of genetic information. A growing number of research results reveal that RNA is important in key processes of regulation and often not translated into proteins [Eddy, 2001, Storz, 2002, Capriotti and Marti-Renom, 2008]. It became clear that RNA is involved in post transcriptional regulation (gene silencing) via microRNAs and small interfering RNAs (siRNA) [Lim et al., 2003, Ender et al., 2008, Xiao and Rajewsky, 2009]. It was also revealed that the translational apparatus is influenced by allosteric conformational changes in riboswitches as well as frameshifts caused by pseudoknots and slippery sequences [Winkler et al., 2002, Penchovsky and Breaker, 2005]. RNA is also involved in the chemical modification of the ribosome [Bekaert et al., 2003] and is even a player in the formation of peptides and, therefore, also important for the production of proteins [Weinger et al., 2004, Nissen et al., 2000]. RNA is important in pathological processes like cancer and retroviral infections such as AIDS [Medzhitov and Littman, 2008]. RNA is also able to form complex 3D structures which are mediated primarily by hydrogen bonds formed between base pairs as well as base stacking because of its single stranded nature. The primary data source for 3D structures of biomolecules is the PDB [Berman et al., 2007]. The number of RNA structures known and being deployed as 3D coordinates in the PDB is growing rapidly. The 3D structure of biochemical elements such as RNAs is often more conserved than its sequence; therefore the structural analysis of RNA becomes increasingly important. In the field of proteins exists a variety of alignment and comparison techniques able to cover a wide range of applications [Kolodny et al., 2005]. In contrast, the field of structural RNA alignment is only now emerging [Dror et al., 2006, Ferrè et al., 2007, Chang et al., 2008].

Opportunities arising from RNA structure string representations

Comparing structures and substructures computationally can be done in multiple ways. Using 3D coordinates of atoms is very precise, but also very restrictive and computationally intensive. In the field of proteins, many methods use the secondary structure as guidance (e.g. [Guerler and Knapp, 2008]). Other methods reduce the chain to elementary characters that can be compared using efficient string matching algorithms [Lo et al., 2007,Gusfield, 1997]. In the field of RNA structural biology approaches are often applied where information of the 3D structure is reduced. Popular methods use dihedral angles of a nucleotide for this reduction (e.g. [Chang et al., 2008]). The RNA society itself released a method that is able to translate an RNA chain into a string using so-called suite codes that are based on nucleotide conformations [Richardson et al., 2008]. Using these reduction methods, it becomes possible to query a huge amount of macromolecules for similarities in a very precise fashion. But in the field of RNA structural biology these fast approaches are not widely used.

Knowledge based docking and protein-protein interaction prediction

A ligand (latin *ligare* - bind to) in the context of biochemistry is a molecule that is able to form a complex with a biomolecule, the receptor. This binding process (“docking”) is reversible and can cause a conformational change of the receptor [Alberts et al., 2002,Lehninger et al., 2008]. A ligand is often a small compound like a drug, peptide or a protein that interacts with another protein. But a ligand can also be DNA bound to protein [Locasale et al., 2009]. The similar property principle is applied to ligands as well as receptors [Barbosa and Horvath, 2004]. A general approach is to search for compounds with a similar structure to find compounds with similar properties. The approach can be extended by looking at the similarity of the binding sites of proteins that might reveal similar binding partners. Consensus exists that the structure of a protein or even a small part of a protein can point towards a specific function. In this regard, many classification databases such as SCOP have been developed [Andreeva et al., 2008]. An example is the Rossman fold that points to a binding of nucleotides and derivates [Rao and Rossmann, 1973]. The Rossman fold highlights that certain conservation among secondary structure elements and residues is important even in the binding of smaller compounds like nucleotides. These small compounds can potentially be drugs affecting a protein. Knowing and predicting the binding partners of a protein can be beneficial, as it may reveal desirable or undesirable effects (ADRs). In drug development ADRs should be omitted wherever possible. In this regard, algorithms that mainly deal with information on an atomic and residual level have been developed [Shulman-Peleg et al., 2008, Yeturu and Chandra, 2008]. Recently, it has also been shown that a certain known interacting protein patch has its application in the domain of protein - protein docking [Günther et al., 2007].

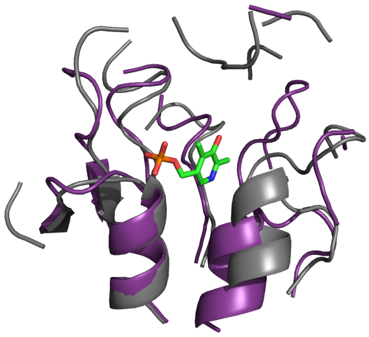


Figure 1.4: Superimposed vitamin B6 binding sites of two proteins (2ctz chain A and 1c4k chain A). The global similarity of the two proteins is low according to SSM [Krissinel and Henrick, 2004] (Q-score: 0.13, % sequence id: 9). However, the local secondary structure of the binding site is highly similar in both proteins.

1.2.3 Cell signaling

Introduction

Section 1.2.2 points out that life is about interaction and regulation. Interaction and regulation occurs everywhere in the cell. There is a growing number of software tools that are able to provide an overview of interactions, most prominently Cytoscape [Yeung et al., 2008, Ruths et al., 2008]. Important databases storing information about interactions on metabolic and signaling level are for instance KEGG and Reactome [Okuda et al., 2008, Vastrik et al., 2007]. Often a simple binary mechanism is anticipated – interaction is happening or not. However, there is almost never a simple on / off switch in nature – it is more about a certain amount of interaction taking place. In this regard, microarray RNA sensing technology has become widely used to predict the amount of protein present in a cell – and in turn the amount of interaction taking place. This is, however, an imprecise view in many cases, as not all RNA is translated into protein. Therefore novel methods are increasingly used that directly measure the amount of a specific protein and the interaction taking place. This finally leads to more and more complete and complex interaction networks [Yu et al., 2008].

Methods

The usage of ordinary differential equations (ODEs) can model interaction between elements of a cell. The ODEs are parameterized according to a set of experiments [Alon, 2006]. When not all interaction parameters are known the view of protein interaction in so-called protein signaling and transduction networks is applicable. They take into account interactions between interaction partners. Signaling commonly has a certain outcome such as "cell proliferation" or "cell death". These signaling networks can be depicted by graphs and frequently the formalism of so-called Petri nets is applied [Grunwald et al., 2008, Heiner et al., 2004]. The methods to analyze a network are commonly divided into static and dynamic properties. Currently, Petri nets are often seen as a static analytical method [Papin et al., 2005] and are not quantified regarding the amount of interaction taking place [Grafahrend-Belau et al., 2008, Chaouiya, 2007].

Pathological cells, perturbation and integration

An important aspect of regulation is how the expression on cellular level is affected by diseases and perturbations of drugs. The NCI (National Cancer Institute) hosts a large scale repository of about 60 cell lines. These cell lines are constantly tested against chemical compounds to evaluate their potential usability as cancer treatment. The NCI measures, for instance, the genetic expression levels of those cells, with the potential to reveal the protein profile of the cell [Ross et al., 2000]. The NCI thus seeks to shed light on the desired and undesired effects of a compound on a cell, even if a compound affects multiple targets [Crespo et al., 2008].

A formalism that is able to represent a signaling network, but that can also be used to integrate a quantification of cellular reactions, would be useful in this regard. The integration of interaction information and perturbation data has the potential to reveal how certain compounds work and which proteins are affected [Chu and Chen, 2008]. It may also be possible to model specific cells or certain organisms and to predict their responses to stimuli and drugs.

1.3 Objectives and overview

The main objectives of this work are:

- Provide solutions for the integration of methods and data (Section 2.1 and Section 2.2)
- Explore novel methods capable of comparing biomolecules and parts of biomolecules in 3D using a string reduction approach (Section 3.1, Section 3.2 and Section 3.3)
- Examine possibilities of using quantitative interaction data (e.g. microarrays) to simulate interactions important in the context of cancer using a graph-based approach (Section 4.1 and Section 4.2)

Chapter 2

Integration of ligand characteristics

2.1 A data warehouse of metabolite and drug binding sites in proteins: SuperSite

The increasing structural information about target-bound compounds provides a rich basis to study the binding mechanisms of metabolites and drugs. SuperSite is a database, which combines the structural information with various tools for the analysis of molecular recognition. The main data is made up of 8,000 metabolites including 1,300 drugs, bound to about 290,000 different receptor binding sites. The analysis tools include features like the highlighting of evolutionary conserved receptor residues, the marking of putative binding pockets and the superposition of different binding sites of the same ligand. User-defined compounds can be edited or uploaded and will be superimposed with the most similar co-crystallized ligand. The user can examine all results online with the molecule viewer Jmol. An implemented search algorithm allows the screening of uploaded proteins, in order to detect potential drug binding sites, which are similar to known binding pockets. The huge data set of target-bound compounds in combination with the provided analysis tools allow to inspect the characteristics of molecular recognition, especially for drug target interactions. SuperSite is publicly available at: http://bioinformatics.charite.de/supersite.

2.1.1 Introduction

The Protein Data Bank [Berman et al., 2007] contains crystallographic information about proteins, which are co-crystallized with thousands of metabolites or drugs. The data is highly relevant not only for analyzing the recognition of individual compounds [Rasmussen et al., 2007], but also as a learning set for molecular interaction models [Schormann et al., 2008]. In many cases, small compounds bound to macromolecules are medically active and listed as approved drugs. The consideration of such co-crystallized structures can considerably facilitate the process of drug development [Bayry et al., 2008]. Another important aspect of molecular interaction is the specificity of a ligand. Many compounds address several receptor proteins. Comparative analysis of the target proteins can enable to draw conclusions about the molecular recognition between ligands and targets [Tikhonova et al., 2008]. One paradigm that frequently reoccurs is the concept of structure activity relationship (SAR) – either meaning, that similar ligands have a similar mode of action [Dunkel et al., 2008], or that similar binding sites may share binding partners. This paradigm has implications for finding novel leads, as well as the elucidation of possible side effects [Minai et al., 2008]. SitesBase [Gold and Jackson, 2006] is an excellent source, which utilizes this similarity concept, using an indexing algorithm that allows for fast comparisons of similar binding sites. This enables the researcher to quickly generate hypotheses about probabilities that a certain binding site will be adopted by a ligand. For further investigations of the interactions between small compounds and macromolecules, a variety of additional sources are available. Concerning experimentally available binding data like Kd, Ki and IC50 data, the Binding MOAD [Benson et al., 2008], PDBbind [Wang et al., 2005] and the Binding Database [Chen et al., 2001] are of special interest, since they allow conclusions about the binding affinity of the compounds. Regarding the integration of secondary databases like SCOP [Andreeva et al., 2004], CATH [Greene et al., 2007] and Pfam [Laskowski et al., 2005], there is a variety of excellent sources with a strong focus on macromolecules, like PDBsum [Laskowski, 2001], RCSB PDB [Berman et al., 2007] and IMB Jena Image Library [Reichert and Suhnel, 2002], while PROCOGNATE [Bashton et al., 2007] is especially tailored for elucidating enzymatic activity. However, there is no single resource, which is centered on drug-like compounds, while integrating all available structural information. Therefore, SuperSite was created with three main design goals in mind:

- Rich integration of the PDB, including full-text search, complete 3D information, and extraction of ligand-receptor relationships.
- Integration of secondary sources, to detect putative binding sites.
- Detection and visualization of compounds considered to be medically active.

The aim of SuperSite therefore is to assist the structural biologist with an online tool, which facilitates the inspection of known and putative binding sites regarding likely binding sites and conservation information. For drug-like compounds, additionally superimposed binding sites of the same ligand are provided, which allows for the detection of structurally conserved residues.

2.1.2 Database and tools

Primary database

SuperSite's main data source is the PDB [Berman et al., 2000], currently containing over 51,000 3D structures and providing well over 290,000 implicit interactions of macromolecules and small compounds. The raw PDB is parsed and translated automatically into a relational database schema that enables SuperSite to further integrate secondary databases for information enrichment (see subsequent subsection). To make the knowledge in the primary database accessible, SuperSite is providing extensible means for querying. The main text query possibilities include the search for PDB-ID, Het-ID, protein, ligand names and synonyms, as well as a full text search, which screens the complete header of all PDB files for a given term. For instance, searching for the term "insulin" reveals all insulin-related proteins so that they can be used for further investigation. An important subgroup of the proteins in the PDB are enzymes involved in many catalytic activities. To this end, SuperSite provides an EC tree presentation [Barrett, 1996] which makes it possible to browse the PDB via enzyme class/subclass and picking proteins of interest. To investigate the similarity of certain proteins, the protein similarity cluster information from the Cd-hit algorithm is integrated [Li et al., 2006]. This information is provided for 95%, 90%, 70% and 50% similarity, based on the sequence. A specialized search form not only allows the search for similar proteins, but also allows for searching apo- and holo-states. This directly allows for dealing with the question, how much the bound form of a protein differs from the unbound form. When it comes to the field of small compounds in the PDB, SuperSite is providing appliances for filtering physio-chemical features like molecular weight, chemical formula or number of atoms. A built-in tool for finding similar small ligands to a given one, is a fingerprint search, based on MyChem fingerprints (http://mychem.sf.net). SuperSite also provides Marvin as an online tool (http://chemaxon.com) which allows to draw or upload a molecule, and screen it against all ligands contained in the PDB (sdf and mol file formats are supported). User-defined compounds are visualized by a superposition according to the most similar bound ligand.

Secondary data sources

To assist the user in investigating potential binding partners and putative binding pockets, SuperSite integrates secondary information from related data sources. Analyses of functionally important sites suggest that the degree of conservation within a protein family is a hint for potential binding sites [Chakrabarti and Lanczycki, 2007]. To this end, SuperSite integrates information from HSSP [Dodge et al., 1998], a data source, which contains information about the degree of residue conservation within a family of proteins. As additional source of information, de novo predictions of possible protein binding pockets are provided that are calculated using LIGSITEcsc [Huang and Schroeder, 2006]. This information is precalculated and also stored in the database. HSSP and LIGSITEcsc provide exhaustive information about putative binding sites. Together with the possibility to elucidate related proteins, this provides starting points for the detection of putative binding sites.

Drug site encyclopedia

A subset of all relations between proteins and small compounds, is the relationship of proteins and drugs. This subset is of high importance when it comes to a systematic investigation of the desired effects of drugs (on- and off-target effects). Therefore, an important part of SuperSite is the Drug Site Encyclopedia. As the term drug is not self-defining, the World Drug Index (http://scientific.thomsonreuters.com), the Comprehensive Medical Chemistry (CMC) Database (http://mdl.com), the NCI cancer compounds (http://dtp.nci.nih.gov) and SuperDrug [Goede et al., 2005a] are compared to all ligands of the PDB to determine the intersection set using standard fingerprints from OpenBabel (http://openbabel.org). The screening was performed via a fingerprint search (http://mychem.sf.net). Entities with a Tanimoto coefficient of > 0.85 and an equal number of nonhydrogen atoms were considered as drugs [Martin et al., 2002]. This screening yielded more than 1,300 medicinal compounds in the PDB. Within the Drug Site Encyclopedia, extended instruments for exploring the relationship between drug and target are provided. One aspect is the possibility to investigate the superimposed binding sites of the same ligand, showing residues that are conserved in a spatial region, or frequently occur in a region characteristic for drug recognition. Additionally, a point set match algorithm is provided, which uses known binding sites (patches) of a ligand, to recognize similar patches on the surface of uploaded structures, solved structures or models (algorithm to be published elsewhere). SuperSite is also calculating Lipinski's Rule Of Five [Lipinski et al., 2001], reflecting the drug-likeness of uploaded or edited compounds.

Visualization, browsing and availability

SuperSite can be used with a standard web browser with active Java 1.5+. The molecular viewer Jmol (http://jmol.org) visualizes proteins, ligands and interactively highlights all integrated data sources like HSSP or LIGSITEcsc. SuperSite also allows for browsing between ligand and protein interactions and vice versa. For instance, it is possible to query the protein "Insulin", pick out a ligand and jump to the next view providing all co-crystallized proteins. If the ligand is contained in the Drug Site Encyclopedia, it is also possible to investigate the superimposed binding sites. Links are provided to numerous relevant sources, containing further specialized data sources (e.g. Proteopedia [Hodis et al., 2008], RCSB [Berman et al., 2000], PDBSum [Laskowski, 2001]). SuperSite is accessible free of charge for academic institutions. Flat files of the database are available upon request.

2.1.3 Case studies

Case study 1: PLP binding partners and spatial mining

Vitamin B6 (Het-ID: PLP) is a co-enzyme, mainly used in the amino acid metabolism and widely present in the human body. Currently, SuperSite contains information about 463 structures containing PLP, representing, a variety of proteins (e.g. aminotransferases, glycogen phosphorylases). A visualization of all binding sites at once can be achieved, by selecting 'Drug Encyclopedia' in the main menu and then entering "PLP" as Het-ID. This view allows inspecting common features, like spatial conservation of specific amino acid types. In the case of PLP, it gets obvious, that, for instance, residue glycine is conserved at a spatial position near the phosphate (Figure 2.1). This is even the case, when the proteins are structurally dissimilar, a conclusion also discussed in [Kume et al., 1991].

Case study 2: Determination of binding pockets

The elucidation of possible binding pockets and active sites of proteins without co-crystallized compounds is a common task for structural biologists. SuperSite provides two tools for the investigation into this topic: LIGSITEcsc – providing precalculated binding pocket predictions and HSSP – providing information about sequence conservation. For instance, PDB-ID 1wdp refers to the structure of the enzyme beta-amylase, solved without substrate. To evaluate if there is a possible binding pocket, the user can consult LIGSITEcsc and HSSP interactively from SuperSite (Figure 2.2). The HSSP conservation shows a more conserved region around residue glutamine (residue number

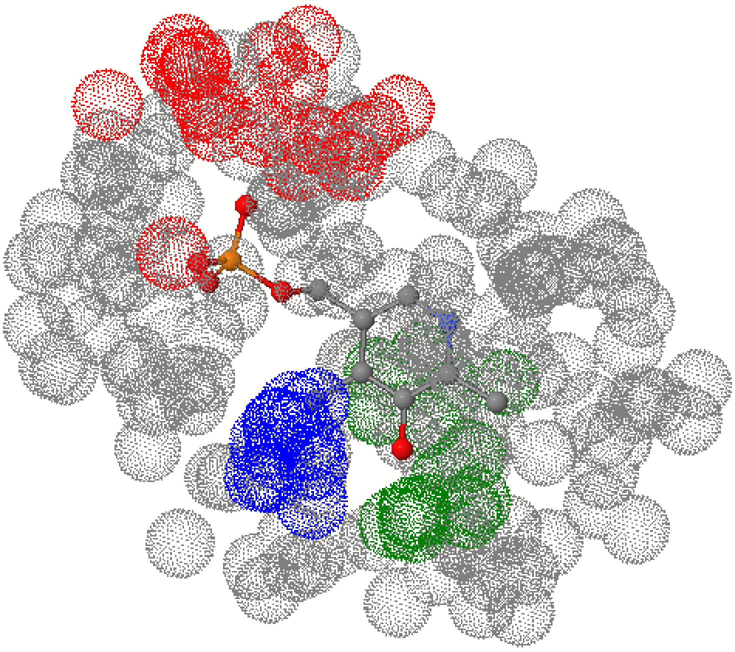


Figure 2.1: Superimposed binding sites of the ligand vitamin B6 (Het-ID: PLP) from PDB-IDs: 1bjo, 1c7n and 1dje. Although the proteins show an overall dissimilar structure, residue glycine (red), lysine (blue) and histidine (green) are clustering at specific spatial positions (other atoms of the binding sites depicted in gray).

186). At the same position, LIGSITEcsc shows a relatively large predicted binding pocket. There is another beta-amylase (PDB-ID: 1b1y) similar in overall structure to the apo form containing a ligand at the position proposed by LIGSITEcsc and HSSP which shows the applicability of this method.

Case study 3: Detection of binding partners via similarity screening

SuperSite also offers a facility for the fast similarity screening of a compound, against all ligands co-crystallized in the PDB. This enables to hypothesize about possible binding partners for similar compounds. Methotrexate (Het-ID: MTX) is a drug, which is used as anti-inflammatory agent/immunosuppressant and in high concentrations used as chemotherapeutical agent [Green and Chamberlain, 2009]. Methotrexate inhibits the folic acid biosynthesis and therefore slows the proliferation of cells. SuperSite enables the user to find similar compounds in the PDB, by simply drawing, or by uploading a mol or

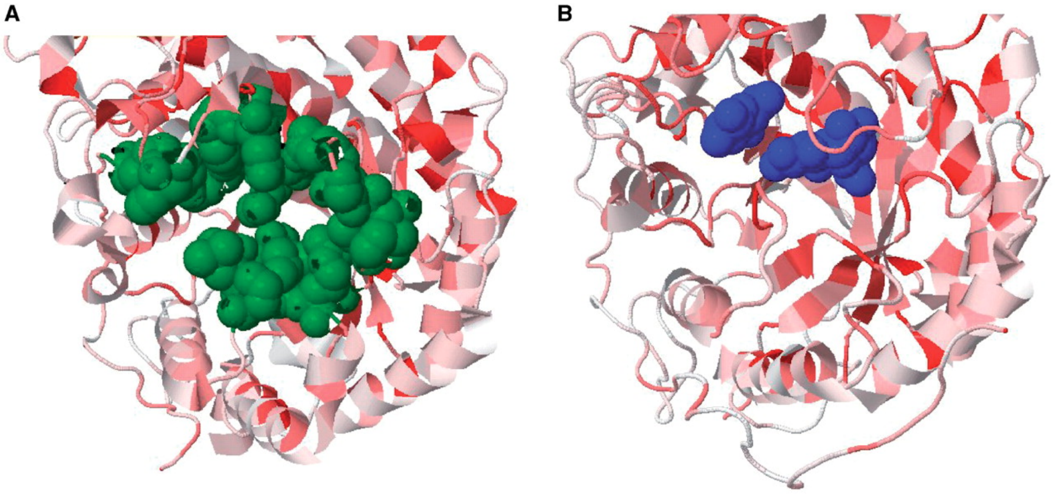


Figure 2.2: An apo (PDB-ID: 1wdp, chain A) and holo form (PDB-ID: 1b1y, chain B) of beta-amylase. The predictions for the binding site pocket (green) as well as the HSSP conservation (red conserved, white not conserved) support the hypothesis of a binding site at this position. This claim can be proved by the holo form (B) with alpha-D-glucose (blue), bound to the predicted pocket.

sdf file. After issuing the similarity search for Methotrexate, one of the best hits not identical to Methotrexate, is folic acid (HET-ID: FOL) bound to a dihydrofolate reductase (Figure 2.3). The query compound Methotrexate is superimposed with folic acid, which is the known mode of action.

2.1.4 Conclusions

SuperSite is a novel database that offers 3D information about proteins and about their bound compounds (ligands). SuperSite enables the user to investigate into the relationship of ligand and receptor in atomic detail, integrating information sources about putative binding sites and conservation on residue level. SuperSite is made with an emphasis on ligands that are drug-like and therefore of special interest for medical research. To this end, SuperSite provides 3D superpositions of all binding sites of a certain ligand, which enable the user to investigate into the spatial arrangement and properties of the binding site. For further investigations, SuperSite allows to issue a similarity screening against ligands bound to macromolecules as well as a screening of proteins against known binding sites.

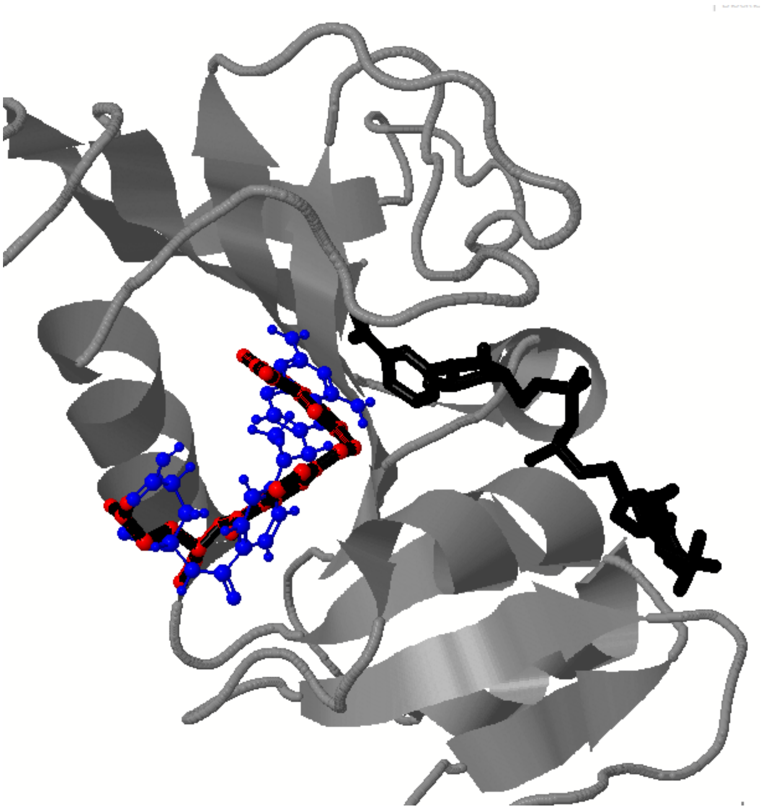


Figure 2.3: A dihydrofolate reductase (PDB-ID: 1ra7) with folic acid (HET-ID: FOL, red) bound. One of the highest ranking results from a ligand similarity screening, using compound Methotrexate (Het ID: MTX, blue), suggests a binding at that position.

2.2 A metaserver for structural search: Superimposé

The Superimposé web server performs structural similarity searches with a preference towards 3D structure based methods. Similarities can be detected between small molecules (e.g. drugs), parts of large structures (e.g. binding sites of proteins) and entire proteins. For this purpose, a number of algorithms were implemented and various databases are provided. Superimposé assists the user regarding the selection of a suitable combination of algorithm and database. After the computation on the server infrastructure, a visual assessment of the results is provided. The structure-based *in silico* screening for similar drug-like compounds enables the detection of scaffold-hoppers with putatively similar effects. The possibility to find similar binding sites can be of special interest in the functional analysis of proteins. The search for structurally similar proteins allows the detection of similar folds with different backbone topology. The Superimposé-server is available at: http://bioinformatics.charite.de/superimpose.

2.2.1 Introduction

As the size of biomolecules differs by orders of magnitude, the ways to compare them and the metrics to measure what a good comparison actually is often differ in the same respect. To cite Hugo Kubinyi: “Similarity lies in the eye of the beholder” [Kubinyi, 1998a, Kubinyi, 1998b]. Therefore, a classification of the alignment problem is required to determine the appropriate method for the detection of the similarity. The definition of similarity in molecular space always depends on the scientific question that is asked. This question heavily influences the design of the algorithm and the definition of the scoring function, which can be adjusted to fit the needs of each request. Unfortunately, comparison algorithms are computationally expensive since the problems are usually NP hard which means that the retrieval of a result is at least extremely time consuming [Lathrop, 1994].

A number of algorithms as well as databases are free for non-commercial use, but in many cases there is no dedicated web server that allows hassle free use of an algorithm and a suitable database to answer a biological question. For small molecules, data sources such as PubChem [Wheeler et al., 2008] and Drugbank [Wishart et al., 2008] provide facilities for similarity searching.

In general, for small molecules their similarity is estimated on the basis of their chemical topology. One method is to translate the chemical topology into so called structural fingerprints. Structural fingerprints are bitvector representations of the small compound chemistry. To compare bitvectors of two molecules metrical coefficients like the Tanimoto coefficent are applied. The Tanimoto coefficient gives values between 1.0 (very similar) and 0.0 (dissimilar). Another often used method is the representation of the molecule as string pattern (SMILES). A simple string search can be used to determine if a certain part of the molecule is present in another molecule or not. But a number of features of small molecules cannot be reflected adequately by 2D representations [Whittle et al., 2003, Chen and Reynolds, 2002]. Recent findings suggest that 3D similarity searches yield at least more varied results [Thimm et al., 2004] than similarity comparisons via the usage of fingerprints or SMILES. Especially to find scaffold hoppers, 3D algorithms clearly show an advantage. For this reason Superimposé is dedicated, but not limited to the usage of 3D algorithms.

There are a number of superposition servers, websites and projects in the field of protein similarity. Often they are merely a companion for a specific algorithm. For instance the website of TM-align [Zhang and Skolnick, 2005] allows to compare protein structures but no search depending on a database. Dedicated superposition servers for proteins include ([Sumathi et al., 2006], [Krissinel and Henrick, 2004], [Maiti et al., 2004], [Leslin et al., 2007] and http://www.ncbi.nlm.nih.gov/Structure/VAST/). 3dSS [Sumathi et al., 2006] has strengths by providing the ability to superimpose more than two proteins. SSM [Krissinel and Henrick, 2004] is a very fast method that even allows searches on a PDB scale level within minutes.

However, due to the fact that algorithms in this field are often domain specific and have their own definitions of good matches, the possibility to choose among a set of algorithms would be beneficial. For a more comprehensive overview about macromolecular superposition the reading of [Novotny et al., 2004, Kolodny et al., 2005] is recommended

For the problem of identifying a similar surface in or on macromolecules there is no website that features such a service for the public yet. Such a service could help to elucidate similar functions of proteins based on shared binding sites or surface patches. Recent findings even suggest that similarities based on interaction patches of proteins can help to get hints about the docking modes between proteins [Günther et al., 2007].

For superposition tasks on Superimposé a three class division of problem cases is defined for molecular similarity searches that branch to different subtasks the user can solve with its help.

- Similarity Class 1: Small molecule level.

- Similarity Class 2: Macromolecule level based on substructures.

- Similarity Class 3: Protein level.

Searches according to Class 1 and Class 3 aim at assigning as many atoms as possible between both structures. For small molecules (compounds) this often means that retrieved compounds are similar in mode of action and/or are affecting similar targets [Barbosa and Horvath, 2004]. Class 2 algorithms are assuming that the query structure is smaller than the macromolecule. A typical scenario for Class 2 algorithms is the identification of similar binding sites. Class 3 specially targets the comparison of entire proteins. The order of amino acids in the peptide chain is valuable information in addition to the 3D-coordinates. In most cases of pairs of homologous proteins the corresponding amino acids appear in the same order. This is because the order of amino acids is preserved in evolution, unless it is disrupted by recombinatorial events leading to circular permutation. However, the number of considered atoms is often reduced by different levels: C-alpha, backbone. Algorithms operating on the protein backbone or even on all-atom-level are often inefficient for protein comparisons [Shakhnovich, 2006]. Established methods therefore often choose hierarchical approaches by dividing the protein into structural elements [Kolbeck et al., 2006].

The preparation of databases, the installation of programs for structure comparison and the sorting and visual inspection of search results is often a complex task with currently available tools. Superimposé facilitates database searches by providing a uniform user interface for different programs, databases and scoring functions. Several databases for small molecules are joined to one comprehensive collection of 3D-structures. Users of Superimposé do not have to solve technical problems and can concentrate on the biological problem.

2.2.2 Algorithms

This alphabetically ordered section gives practical descriptions of algorithms deployed by Superimposé. If not stated otherwise Superimposé uses original binaries with default parameters for the algorithms.

GangstaLite

GANGSTA [Kolbeck et al., 2006] is an algorithm for structural alignment of proteins and similarity search. GangstaLite is a specially drafted fast version for the Superimposé project. GANGSTA works in two stages: In the first stage, a mapping on the secondary structure elements is generated using a combinatorial approach that replaces the former genetic algorithm. In the second stage, individual residue pairs are assigned to create a maximum contact overlap.

GangstaLite is designed to detect similarities between proteins without using sequential information. Therefore cases of fold similarity without sequential similarity will be recognized. An example of circular permutation is presented in the case studies.

NeedleHaystack

NeedleHaystack [Hoppe and Frömmel, 2003] computes structural alignments of molecules as superpositions of sets of single atoms in the 3D-space where information on chemical connectivity and atom types is not necessarily considered. It is specially suited to scan a large molecule (target = haystack, up to 100,000 atoms) for the occurrence of a given molecular motif (model = needle) with a given tolerance level. It operates on the complete enumeration of superpositions of atom triples in both model and target but radical pruning reduces the running time to seconds for a typical problem size, the search for a binding site in a protein surface. As NeedleHaystack is used for binding site recognition the parameters -sk 0.25, -ad 1.35, -al 2, -to 60, -bd 1 are applied. Additionally, NeedleHaystack uses a weighting matrix that punishes each missed superposition on atom level with the score 2.

A typical application for this algorithm is the search for similar binding sites. This is illustrated in the case studies section.

PSM

PSM [Formella, 2005] is a program that finds and aligns a small search pattern in a large search space, e.g., some sort of known substructure in a possibly large protein. PSM is an efficient implementation of a sub graph matching algorithm that uses certain domain specific heuristics. The atoms represent the vertices of the distance graphs; their distances among each other represent the edges of the graph. The lengths of the edges of the distance graph over the search pattern are used to construct the distance graph over the search space where only the edges that have similar lengths as the corresponding edges in the search pattern are maintained.

With the help of a backtracking algorithm, PSM enumerates all possible matches. Heuristics are used to order the vertices and edges during the search in such a way that the algorithm discards non-profitable partial matches early. The heuristics include, for instance, atom type, membership to a certain chemical group of the atoms, frequency of edge distance in the graphs. PSM not only finds the ideal alignment based on dRMS (distance root mean square), but is able to compute the (locally) optimal alignment for average distance, maximum distance or any other distance metrics. PSM uses the derivative free minimization algorithms taken from [García and Rodríguez, 2002] to compute the rigid motion transformation, including a small scaling factor. Due to the fact that PSM is based on distance graphs, it can be easily extended to work with deformable search patterns where hinges and torsions are allowed. Furthermore, individual tolerances can be assigned to all edges and L-matches (i.e., mirrored matches) can be found.

PSM is able to recognize similar surface patches / active sites.

Score1

For the scoring of a partial superposition M (i.e., partial matching of atoms) between the two input molecules, Score1 applies *score* of M (Definition 2.2.2), based on the RMSD (Definition 2.2.1).

Definition 2.2.1

$$\mathrm{RMSD} = \sqrt{\frac{1}{n}\sum_{i=1}^{n} d_i^2} \tag{2.1}$$

where d_i is the Euclidian distance between N pairs of equivalent residues. The RMSD is calculated in Ångström.

Definition 2.2.2

$$score(M) = r \cdot \exp(-RMSD(M)), \tag{2.2}$$

where r is the proportion of superimposed non-hydrogen atoms of the smaller molecule and $RMSD(M)$ is the square root of the least possible mean squared distance between atom pairs matched in M under all possible rigid motions of the input molecules.

Therefore $score \in (0.0, 1.0]$ acts as a geometric similarity measure between two input molecules. If one molecule is identical to another molecule, then there is a superposition M such that $score(M) = 1.0$. Score1 calculates an optimal spatial superposition of two drug-sized molecules with respect to the above score function subject to an additional constraint: For every atom a matched in the superposition, there has to be an atom b bound to a such that b is matched, too. This restriction of the search space allows using an optimal branch-and-bound algorithm as described in [Thimm et al., 2004] without any reduction of the input molecules. To speed up the algorithm, also lower bounds for possible solutions along different paths in the search tree are calculated. Promising paths can be searched first, leading to a more effective pruning. To establish the lower bounds, techniques from [Kirchner, 2007] for calculating the optimum atom pairs given a fixed rigid motion of the input molecules are used. In accord with the authors the parameters “0.7 0.65 0.0” are used to enable Score1 being used in whole database screening applications.

Score1 is suitable for similarity screening in small molecule databases, illustrated in the case studies section.

sd_best_compare

The algorithm sd_best_compare is based on a normalization of the atomic sets according to their principal moments of inertia [Preissner et al., 1999]. This first normalization is of course independent of transformations of the coordinate system, and quite stable for small alterations of the atomic positions. It is also unique except for four possible rotations. Therefore, the degree of freedom is strongly reduced, and the assignment of pairs of related atoms is straightforward for identical or very similar sets. In a first step both atomic sets are roughly orientated according to their size proportions. After superimposing the centers of mass and alignment of the longest and smallest dimensions closest atoms are assigned as pairs. This assignment is improved by numerous refinement cycles. The algorithm was tailored for the search of similar atomic sets in a large data base of patches (not necessarily bonded atoms) [Frömmel et al., 2003]; the aim of the algorithm is not to compare very different molecules but to find similar molecules with different connection schema. To do this as fast as possible the database should be prepared to minimize the effort of parsing the data file [Preissner et al., 2001]. With the help of some adapted procedures the method can also be used to compare entire proteins.

The algorithm was implemented to compare conformational databases of low molecular weight structures that share similar scaffold [Thimm et al., 2004].

TM-align

TM-align [Zhang and Skolnick, 2005] uses a two step process that is made up of an initial structural alignment based on an initial assignment of SSEs and dynamic programming. This step is followed by a heuristic optimization. The alignment as well as the heuristic optimization is based on TM-score. TM-score is a variation of the Levitt-Gerstein weight factor that punished larger distances relatively stronger than smaller distances and allows more sensitivity concerning the global topology. The value of TM-score lies in (0,1]. In general, a comparison of score < 0.2 indicates that there is no similarity between two structures; generally, a TM-score > 0.5 indicates that structures share the same fold, but the drop-off of the score indicating the twilight-zone of similarity has to be considered individually.

TM-align is an algorithm for protein structure alignment.

CE (Combinatorial Extension)

The algorithm CE [Shindyalov and Bourne, 1998] involves a combinatorial extension of an alignment path defined by aligned fragment pairs (AFP), which represent possible alignment paths. Combinations of AFPs are selectively extended or discarded to yield an optimal alignment path. They are based on local geometry, rather than global features such as orientation of secondary structures and overall topology. The algorithm is fast and accurate in elucidating structural alignments and fast enough for database scanning and detailed analyses of protein families.

CE builds an alignment between two protein structures.

2.2.3 Databases

This section provides information about the databases in alphabetical order. Databases are updated on a monthly basis.

Astral 40

The Astral Compendium [Chandonia et al., 2004] provides several databases and tools derived partly from the SCOP [Lo Conte et al., 2002] database and based on PDB coordinate files. SCOP itself provides schemes of all proteins available in the PDB according to their evolutionary and structural relationships. Additionally, a grouping of proteins into species and a classification into families and superfamilies, folds and classes is provided. ASTRAL 40 provides this information filtered with 40% sequence identity in a PDB style format that is deployed onto the Superimposé web server. Astral provides 9,500+ chains / domains and aims to represent the whole structural space of proteins. A link to the PDBSum [Laskowski, 2007] is provided that enables the user to examine the found proteins in great detail with the original paper.

Ligand Depot

The Ligand Depot [Feng et al., 2004] is a data warehouse that integrates databases, services, tools and methods related to small molecules bound to macromolecules. It provides chemical and structural information about small molecules in entries of the Protein Data Bank. Currently, it contains information about 80,000+ structures. All small structures of the Ligand Depot are deployed on the Superimposé server and allow to search for the occurrence of small molecules or analogues in the PDB.

Open NCI Database

The release of the Open NCI Database [Voigt et al., 2001] includes 210,000+ compounds with 25 conformers on average. The Open NCI database contains compounds that show a significant activity as therapeutic agent against diseases like AIDS and cancer. A molecule that is highly similar to a compound in the Open NCI might have similar medical activities. For further investigation a link to the Enhanced NCI Database Browser [Ihlenfeldt et al., 2002] is provided.

PDB (Culled)

The PDB [Berman et al., 2000] is an archive of experimentally-determined, biological macromolecule 3-D structures and contains 48,500+ structures of proteins. Because of the nature of the PDB as all purpose repository for macromolecules it often contains duplicate structures and structures of a resolution that are hardly suitable for searching. Another problem is the sheer size of the PDB, what makes it impossible for many algorithms to perform comparisons between proteins (Class 3) and on substructures of proteins (Class 2). For both reasons a representative subset of the PDB is used. The subset is calculated using the PISCES Server [Wang and Dunbrack, 2005]. The used cut-off thresholds are: Sequence identity cut-off: 20%; Resolution cut-off: 1.8Å; R-factor cut-off: 0.25. A link to the PDBSum is provided.

PDB Surfaces (Culled)

For the elucidation of similar parts on the surfaces of macromolecules it is suitable to limit the search space to the water accessible surface. None of the presented algorithms does this on its own, so a precomputing step is applied for the PDB (Culled) Database described above. The algorithm calc-surface [Tsai et al., 1999] is used to generate macromolecules with the water accessible surface alone. A link to the PDBSum is provided.

Superdrug

The Superdrug [Goede et al., 2005b] database contains 2,500+ 3D-structures of active ingredients of essential marketed drugs. To account for structural flexibility they are represented on average by about 40 structural conformers per drug generated by the program Catalyst (Accelrys Inc. http://www.accelrys.com). Superdrug provides a link to the Superdrug website that enables the user to investigate results in more detail like the ATC code (WHO

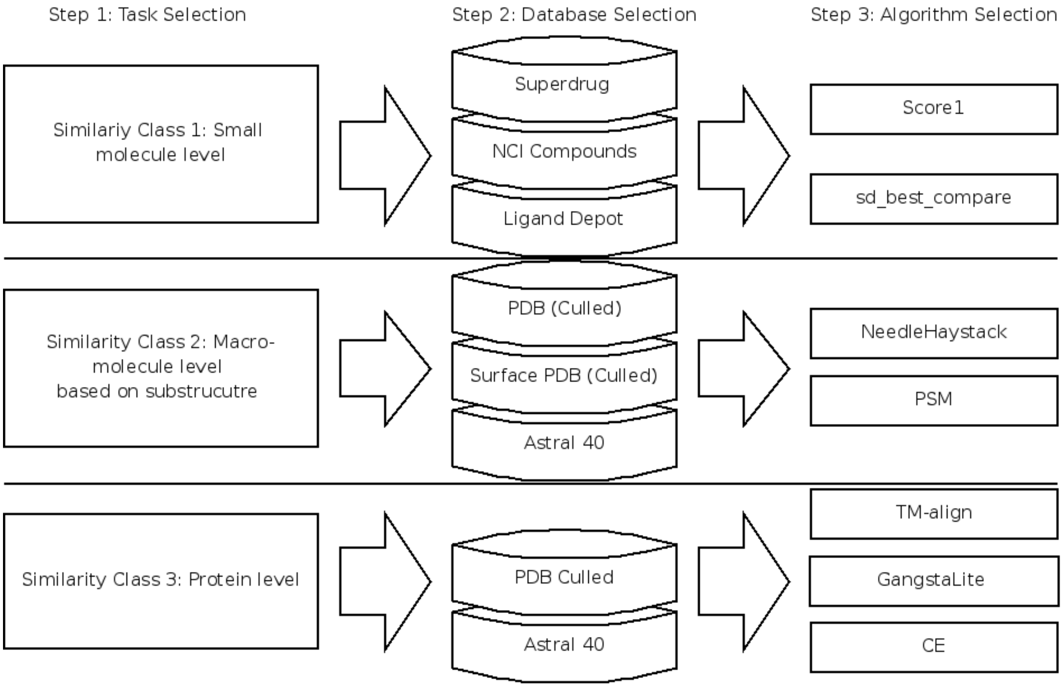

Step 1: Task Selection
Step 2: Database Selection
Step 3: Algorithm Selection
Similariy Class 1: Small molecule level
Superdrug
NCI Compounds
Ligand Depot
Score1
sd_best_compare
Similarity Class 2: Macro-molecule level based on substrucutre
PDB (Culled)
Surface PDB (Culled)
Astral 40
NeedleHaystack
PSM
Similarity Class 3: Protein level
PDB Culled
Astral 40
TM-align
GangstaLite
CE

Figure 2.4: Suitable combinations of databases with algorithms depending on the class of the scientific problem.

classification of medical compounds according to their therapeutic application and chemical scaffold).

2.2.4 Web server description

For Superimposé it has been decided to provide a wizard style approach that guides the user through the different possibilities on offer (Figure 2.4). A fixed set of parameters for all algorithms is used that allow a generalized execution of tasks. A typical search workflow begins with the selection of a task the user wants to execute. This tasks maps to the three classes described in the introduction. In a next step, the user can upload a file to act as model (or patch in Class 2) for the search. Supported file formats are sdf, mol and pdb. Conversions between different file formats are handled via OpenBabel [Guha et al., 2006]. Subsequently, the user gets a selection of suitable databases and algorithms for that task.

Computations can take longer times (24h) in case there are several users employing the web service. Therefore, the user provides an email address, where a report about finished jobs is directed to. This email contains a hyperlink to a webpage on the Superimposé server that presents all results for the search with possibilities to visually assess the results. A specially designed visualization via Jmol as a Java Applet is provided. This allows the user to execute custom scripts in the Jmol language for extensive visualization. The second visualization possibility especially tailored for proteins is STRAP [Gille and Frömmel, 2001] which is implemented via Java Webstart and behaves like a native application and not like a webpage as Jmol does. For both programs the sole requirement is a Java JRE (http://java.com).

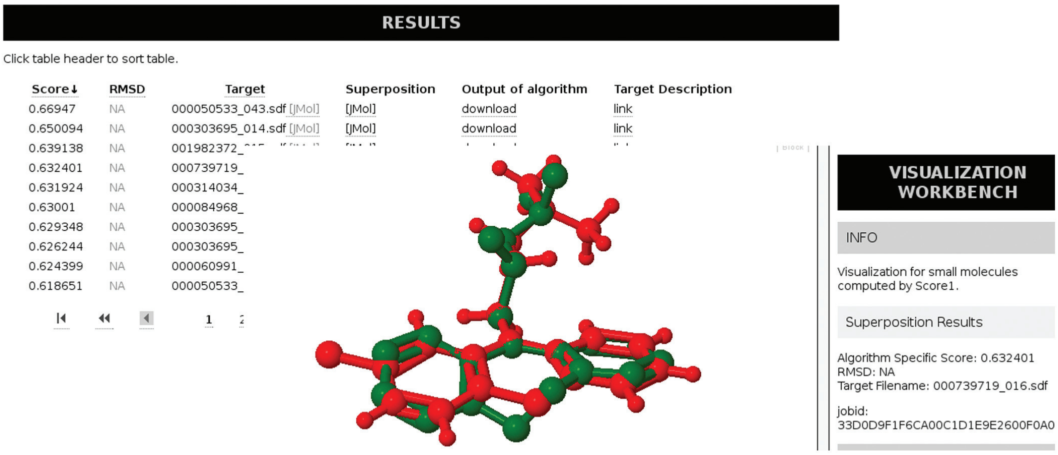
2.2.5 Case studies

The following case studies are organized per problem class and show typical problems where Superimposé can be applied. All molecules and proteins that are discussed within the case studies are available for download on the Superimposé web page (documentation).

Small structure similarity (Class 1)

Similar compounds are more or less likely to share properties such as ligand specificity and binding strength. Thus, screening for similar compounds in databases is a standard technique to generate new hypotheses for molecules (shared activity). Therefore, Superimposé allows the user to search for similarities against a variety of compound databases. In this case the ability of Superimposé to successfully retrieve similar compounds to Chlorpromazine (ATC: N05AA01) on the database Superdrug with the algorithm Score1 is highlighted. Similarity is defined as the ability to find compounds in a related ATC group. The results for the first 10 entries show that Superimposé is able to find compounds that are apart from two compounds Methdilazine (ATC: R06AD04) and Pimethixene (ATC: R06AX23), all coming from the desired ATC-code **N** (**N**ervous System). For the two compounds from ATC group **R** (**R**espiratory System) this could point to unwanted side-effects of Chlorpromazine. The fingerprint-based search on the website of Superdrug fails in retrieving the compounds Trimipramine (ATC: N06AA06) and Cyamemazine (ATC: N05AA06).

Compared with the results of the Superdrug website Superimposé is additionally able to successfully retrieve compounds Trimipramine (ATC-code N06AA06) and Cyamemazine (ATC-code N05AA06) which are left out by the fingerprint search. The reason is that structural superposition is able to superimpose scaffold hoppers, in this case a six- and seven membered ring structure (Figure 2.5), which are dissimilar in the SuperDrug fingerprint search.



RESULTS

Click table header to sort table.

Score↓	RMSD	Target	Superposition	Output of algorithm	Target Description
0.66947	NA	000050533_043.sdf [JMol]	[JMol]	download	link
0.650094	NA	000303695_014.sdf [JMol]	[JMol]	download	link
0.639138	NA	001982372_			
0.632401	NA	000739719_			
0.631924	NA	000314034_			
0.63001	NA	000084968_			
0.629348	NA	000303695_			
0.626244	NA	000303695_			
0.624399	NA	000060991_			
0.618651	NA	000050533_			

VISUALIZATION WORKBENCH

INFO

Visualization for small molecules computed by Score1.

Superposition Results

Algorithm Specific Score: 0.632401
RMSD: NA
Target Filename: 000739719_016.sdf

jobid:
33D0D9F1F6CA00C1D1E9E2600F0A0

Figure 2.5: Query compound Chlorpromazine (red) and search hit Trimipramine (green).

Substructure search (Class 2)

Here, the ability of the NeedleHaystack algorithm together with the Culled-PDB is shown to identify related proteins based on a patch from the catalytic site. For the case study, a patch from the active site of protein Hydrolase (PDB-code: 1pek) is used. This patch is successfully identified on a Subtilisin complex (PDB-code: 2sic) with related activity. NeedleHaystack retrieves perfect matches e.g. in the active site of 2sic (Figure 2.6).

Protein similarity (Class 3)

For the problem of protein similarity/protein alignment a main case where sequence-based methods often fail is for proteins that are similar in terms of overall structure (fold) but not on sequence level. One example where especially the GangstaLite algorithm can find meaningful alignments is an Integrin alpha-V (PDB-code: 1m1x). In combination with the Astral database GangstaLite successfully retrieves a WD40 domain of the Transcriptional Repressor TUP1 (PDB-code: 1erj) as one of the best scoring alignments (Figure 2.7). GangstaLite successfully aligns the proteins with half of the secondary elements not in sequence direction.

2.2.6 Conclusions

Superimposé is created to deal with structural superpositions of molecules in a widespread sense. The combination of databases and algorithms of different

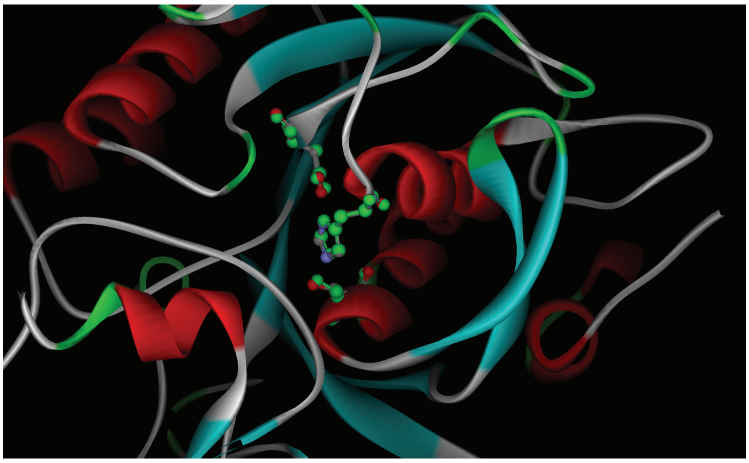
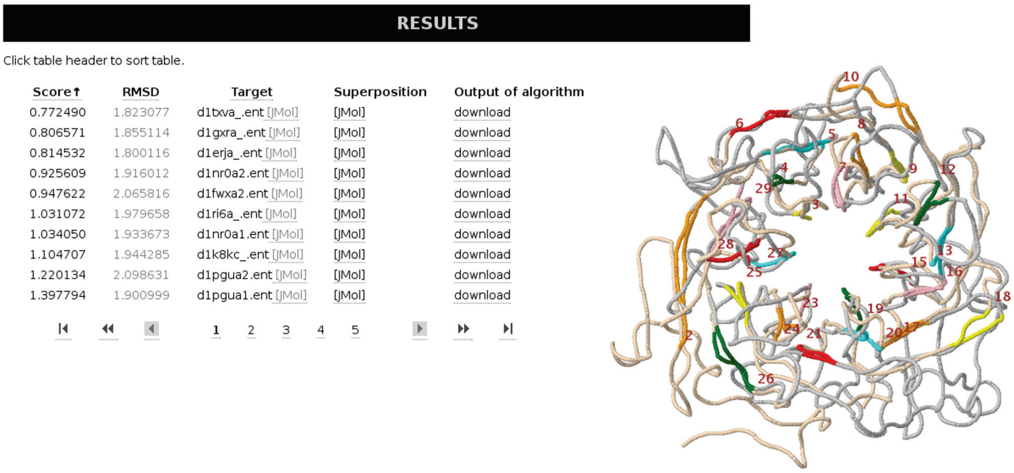


Figure 2.6: Superposition of the active site derived from protein Hydrolase (PDB-code: 1pek / green atoms) that is successfully identified in protein Subtilisin (PDB-code: 2sic / cpk colored ball-and-sticks in the middle).

fields provides amongst others the possibility to identify similar proteins, similar medical active compounds and also binding-sites via similarities in substructure search. The server will be useful for bioinformaticians specialized on structures, macromolecular biologists and the systems biology community by providing possibilities to identify similar patches (binding sites / surface patches) in known proteins. By reducing the complexity of installing algorithms, databases and finding suitable parameter sets Superimposé allows researchers to instantly deal with the task without the administrative problems around it.



RESULTS

Click table header to sort table.

Score↑	RMSD	Target	Superposition	Output of algorithm
0.772490	1.823077	d1txva_.ent [JMol]	[JMol]	download
0.806571	1.855114	d1gxra_.ent [JMol]	[JMol]	download
0.814532	1.800116	d1erja_.ent [JMol]	[JMol]	download
0.925609	1.916012	d1nr0a2.ent [JMol]	[JMol]	download
0.947622	2.065816	d1fwxa2.ent [JMol]	[JMol]	download
1.031072	1.979658	d1ri6a_.ent [JMol]	[JMol]	download
1.034050	1.933673	d1nr0a1.ent [JMol]	[JMol]	download
1.104707	1.944285	d1k8kc_.ent [JMol]	[JMol]	download
1.220134	2.098631	d1pgua2.ent [JMol]	[JMol]	download
1.397794	1.900999	d1pgua1.ent [JMol]	[JMol]	download

Figure 2.7: Results (left) and non-sequential structural alignment generated by GangstaLite (right).

Chapter 3

Biomolecular search, similarity and coding properties

3.1 RNA character encoding and suffix techniques

The RNA Ontology Consortium recently proposed a two-letter representation of the RNA backbone conformation. In this study, the suite notation is compared to a custom string representation that utilizes $\eta - \theta$ pseudo torsion angles. Both representations were used to assess similarity and self-similarity in several RNA structure datasets. For the detection of similarities between two RNA structures suffix techniques are utilized that allow for the detection of substructure similarity within some degree of inexactness. The suite representation as well as the pseudo torsion representation was tested on four diverse RNA datasets. The possibility to detect structural similarities on these datasets allowed recovering many similar structural elements that have implications for further understanding of the RNA apparatus in systems biology. The software as well as the utilized datasets are freely available from http://suiterna.sourceforge.net.

3.1.1 Introduction

String-based approaches to RNA structure analysis are widely used as long as secondary structures are concerned. But, there have been few attempts to express 3D features in a string notation. Recently, the RNA Ontology Consortium [Leontis et al., 2006] proposed a string representation for the conformation of RNA backbones. This allows the use of classical string matching methodology to compare structural features in turn. This work explores how suffix techniques can be used to find similar regions in RNA backbone strings.

RNA secondary structures are most commonly expressed in the dot-bracket grammar, which contains all nested Watson-Crick and wobble base pairs. This string notation is easy to handle, and therefore has been widely used to describe local motifs [Hofacker et al., 2004], for computational approaches comparing RNA sequences by tree grammars [Reeder et al., 2006], and for aligning two or more sequences [Dowell and Eddy, 2006]. To distinguish subtle structural motifs, like the sarcin-ricin motif, RNAse P, pseudoknots, and tertiary interactions, this notation is not enough. These features depend on specific base pairing and stacking interactions, and a specific arrangement of the RNA backbone.

The RNA Ontology Consortium has bundled efforts to describe RNA structures. It poses a platform where structural Bioinformaticians can exchange ideas and discuss formal nomenclature. Systematic approaches to describe RNA tertiary structure have been started from many sides: A typology of base pairs as the basic unit of which RNA is built was defined [Yang et al., 2003]. This allowed to identify interchangeable pairs of base-base interactions (known as the isostericity principle) [Lescoute et al., 2005]. Stacking is conceived as a major stabilizing force, and two complementary typologies have been introduced [Lescoute and Westhof, 2006]. To describe larger local structural units, circular topologies, residues interconnected by backbone, base-pair or stacking interactions, have been introduced. Assembly of these building blocks has been successfully used in constructing tertiary structures, given that the topology is known or well-predicted [Parisien and Major, 2008].

Richardson et al. created a string representation of the RNA backbone [Richardson et al., 2008], where the backbone conformation of ribose-to-ribose "suite" units can be represented by two letters. To analyze the RNA backbone, the most significant features are torsion angles. For each base, there are six of them, one for each bond from one phosphodiester unit to the next. These torsion angles show a characteristic distribution. More distinct clusters of the torsions can be found if RNA 'suites' – units from one ribose to another – instead of the traditional phosphate-phosphate units are considered [Murray et al., 2005]. Each suite consists of seven torsion angles, including both C4'-C3' bonds. The torsion angles were clustered, each cluster being defined as a hyperellipsoid in the 7D space formed by the seven torsions of one suite. In total 46 distinct conformations of the backbone were identified. For each cluster, a two-character code was assigned. The first character corresponds to the first three torsion angles, and the second to the other four. Thus, it is possible to write an entire RNA 3D structure as a 1D string representing the backbone.

The main disadvantage of the suite representation is that its scope is limited to well-defined backbones. For a high quality dataset, it covers 90%-95% of the residues in RNA structures. The other residues are disregarded either because any of the backbone torsions are outside well-defined boundaries, or because the suite is not close enough to any of the hyperellipsoids in 7D space. Most of the unassigned residues are in flexible regions having a high temperature factor, or they simply belong to clusters that are too sparsely populated to form a separate cluster.

An alternative description of the RNA backbone is based on pseudo torsion angles. For this, the RNA structure is reduced to C4' and P atoms – similar to the Cα trace of proteins. Between these atoms, two pseudo torsions η and θ are defined. Even though it is more coarse-grained, the $\eta - \theta$ angles encode important features such as the sugar pucker to a satisfying degree. The Amigos program can be used to calculate pseudo torsions [Duarte and Pyle, 1998]. The P and C4' atoms are frequently used to construct initial backbone trace in X-Ray crystallography. Recently, it was reported that using P-C1' pseudo torsions improves the assignment of the backbone and ribose to electron density maps (K. Keating, personal communication), but it was not explored how these pseudo torsions map to other structural features.

It is very tempting to utilize these backbone representations to compare local structures of RNA to each other. There are only few instruments available to compare RNA structures. Most of them are based on secondary structures, and they use the dot-bracket grammar. Among them, RNAforester [Reeder et al., 2006], Vienna [Hofacker, 2003] and ARTS [Dror et al., 2006] are the most common. Recently a web server (SARSA) was released [Chang et al., 2008] that uses a custom vector quantification to cluster the RNA bases into 23 distinct conformers that are translated into a string representation. SARSA is subsequently applying traditional string alignments to find similar motifs. SARSA is especially useful when applied to multiple alignments of RNA structures; however a search against a database of RNA structures is not supported. The RNAFRABASE web site (http://rnafrabase.ibch.poznan.pl/) contains a big number of loop fragments from RNA structures, but it is very limited in both the kind of fragments contained, and possible search methodology.

Currently, there exists no method that allows fast queries for similar RNA substructures against a database. Therefore, it was decided to use string representations of the RNA backbone in order to take advantage of existing algorithmic solutions for the efficient string search. Alternatively a pseudo torsion representation of $\eta - \theta$ angles is calculated. To cope with the problem of thousands of motifs and thousands of RNA structures available a suffix technique [Giegerich and Kurtz, 1997] is used that holds all information in an index and can be crawled almost linearly.

The main objectives are in brief:

1. Verification of the applicability of the RNA Ontology Consortium suite code, by examining the suites of differently structured RNA.

2. Presentation of a suffix method to compare RNAs to each other and giving an overview which structures and substructures are similar.

3. Discussion of possible alternatives (regarding the structure - string coding, used search algorithms) and applications.

3.1.2 Methods

Suffix arrays are constructed from strings consisting of the RNA Ontology Consortium suite codes for four different datasets: motifs from the SCOR database, all tRNA structures, a high-resolution dataset, and the representative RNADB05 set. Each of them was then queried for matching subsequences in the suffix array to detect structural similarities. As an alternative approach, strings representing $\eta - \theta$ angles of the RNA backbone were constructed and processed in the same way.

Datasets used

SCOR dataset Of foremost importance was to know, whether known RNA motifs annotated in SCOR can be recovered by the suite representation. SCOR is a database containing 15,945 structural, functional and tertiary interaction motifs that have been annotated manually [Tamura et al., 2004]. A hierarchical classification inspired by the SCOP database [Andreeva et al., 2008] has been established, but the database lacks updates after 2004. Therefore, a reliable automatic recognition of motifs could be useful. Currently, no such procedure is available with the circular motif library of the MC-Sym program probably coming closest [Parisien and Major, 2008]. For this analysis, all 4,501 structural and 100 tertiary interaction motifs from SCOR (version 2.0.4) data were used. Functional motifs annotate entire RNAs, and were excluded. The according fragments of PDB structures had lengths between 2-11 suites for structural, and 4-60 suites for tertiary interaction motifs. This set was termed “SCOR”. Functional motifs are annotating entire RNAs, and are considered in the later datasets.

tRNA dataset An interesting aspect is whether structurally highly conserved RNAs can be recognized by the suite representation as a positive control. For this, the tRNA as one of the most conserved molecules in life was chosen. Although tRNA sequences started diverging even before the genetic code itself was fixed and their structures are highly modified by post-transcriptional additions, all of them need to have a highly conserved tertiary structure in order to work in the translation machinery. Thus, it is not surprising, that all example tRNAs from the PDB look the same from afar – and it can be expected that they should have very similar backbone conformation when represented as suites. To examine whether this hypothesis holds, all tRNA structures from the NDB database [Berman et al., 2002] were retrieved. The resulting tRNA set consists of 102 tRNA structures from all kingdoms of life and is termed "TRNA".

RNADB05 and HIRES sets Another objective was to check for similarities among RNAs of different origin. This was done for two sets of RNA structures. One was the dataset used by Richardson et al. (termed RNADB05) [Richardson et al., 2008]. The RNADB05 set is a manually refined representative set of 173 RNA structures from both X-Ray and NMR experiments. The second set (HIRES) consists of 74 high-resolution X-Ray structures. They were filtered from the PDB by applying resolution <= 2.5 Å and r-value <= 0.25 constraints. Structures with identical sequences, and sequences with less than four bases were discarded.

Calculation of RNA backbone string representation

For each structure in each of these datasets, a string using the suite representation, and another one based on the pseudo torsions was calculated. The calculation is also applied to structures that are queried against one of these datasets.

The method to calculate suites from a structure was re-implemented according to the description in [Richardson et al., 2008]. The seven torsion angles were calculated according to Figure 3.1 in 5' to 3' direction. They were then assigned to one or none out of the 46 suite clusters. First they are grouped according to their δ, $\delta - 1$, and γ angles to limit the number of clusters to be considered. Second, the 7D distances to the 7D hyperellipsoids for each cluster were calculated. If the suite was inside a hyperellipsoid, its name was assigned to the suite. The extent of these hyperellipsoids varies depending on the cluster. Especially, some of the clusters were partially overlapping; in these cases the closest hyperellipsoid center was used.

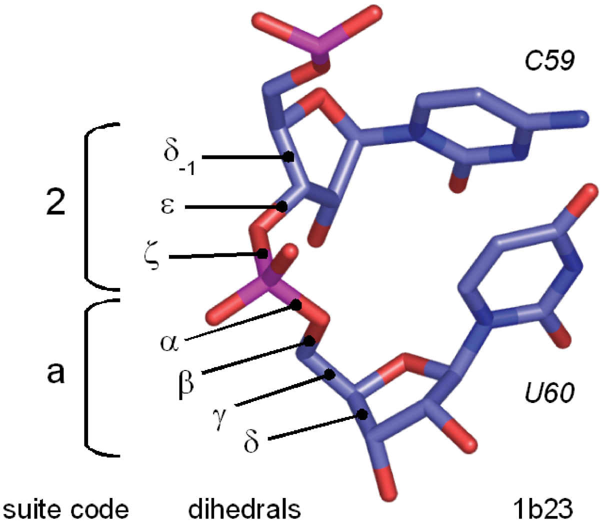


Figure 3.1: Definition of RNA suites. A suite stretches from one ribose unit to the next, involving seven dihedral angles along the RNA backbone. Note that the δ angle is used by two adjacent suites. In the suite encoding, the first three dihedral angles are represented by a number, the next four by a letter. The example is taken from the tRNA structure with PDB-code 1b23.

Even though it is recommended by Richardson et al. not to calculate suites for residues with a high B-factor and with clashes, it was decided to include them anyway. This was done for two reasons: First, to have a continuous string representation for all RNA structures. This is particularly important considering that 5-15% of the residues are not assignable to suites, and thus in average only short fragments of structure would remain for calculation at all.

Second aspect was the number of errors that occur in a real-life dataset. There were four kinds of errors: Missing atoms in the residue (resulting in a '−−' suite code), a single torsion angle outside boundaries defined in [Richardson et al., 2008] (so-called triaged residue, resulting in a 'tt' suite code), an outlier suite which is not close to any cluster (resulting in a 'oo' suite code), and a close outlier inside a 4D hyperellipsoid but outside in 7D space (resulting in a '!!' suite code).

The second possibility to translate a 3D structure of an RNA into a sequence of characters is implemented by calculating the $\eta - \theta$ pseudo torsion angles from the backbone atoms of the same residues as the suites. For η, these were the C4'i-Pi+1-C4'i+1-Pi+2 dihedral, and for θ the Pi-C4'i-Pi+1-C4'i+1 dihedral angles. Each of these angles was divided into 36 ten-degree bins, and for each bin, an alphanumeric character was assigned. Thus, a single $\eta - \theta$ tuple – conceptually corresponding to the RNA suite – was represented by two characters as well. Only in the case when either of the atoms defining the dihedral was missing, an '−−' code was assigned in place of the $\eta - \theta$ tuple.

Suffix tree and array implementation

The studies where performed using a suffix array. While even simple implementations of suffix trees fulfill the property to search for a given substring in O(m) with m being the length of the input string it was decided to use the slightly slower suffix array implementation because of a better memory footprint. An algorithmic introduction to suffix trees and suffix arrays is given in [Gusfield, 1997]. The implementation used as suffix array can search in O($m$log$n$) with m being the length of the search string, and n the number of strings in the index. This performance is fast enough considering the absolute amount of structures to index – even for all RNA structures in the PDB (currently 1500).

A suffix array works in principle in the following manner: To index a string s with length m in the suffix array each substring from $0 - m$ is put into an array. This array is then sorted alphabetically. After the sorted array is established a substring of s can be retrieved by using binary search over the index that fulfills the O($m$log$n$) property.

A conceptual disadvantage of the suffix techniques applied is that a substring search can only be performed in an exact manner. To overcome this disadvantage the notion of n-grams is applied to perform an inexact search and to get a scoring of one input structure against a whole database. This similarity score (*SCORE*) is generated by searching all consecutive substrings of length n (n-grams) of the input string against the database.

$$SCORE = \frac{number\_of\_matches\_found}{number\_of\_matches\_expected} \tag{3.1}$$

This allows for a ranking of the best matching entries in the database as well for a nice way to generate an all-against-all ranking of entities in a database. One drawback of this scoring scheme is that ubiquitous repeating substrings (like '1a1a1a1a') are found in nearly every entity in the database and therefore add a huge bias to the calculation. To avoid that, a search of substrings with repeating entities is excluded.

Apart from the theoretical runtimes given by $O(x)$ the practical runtimes for the n-gram search with the current Suffix Array implementation is below 5 seconds for an all against all search of the RNADB05 set (257 entries) on a commodity pc (dual core 2.2 GHz, 3 GB RAM).

3.1.3 Results

In this analysis, it was systematically looked for similar backbone conformations, and then checked whether they occur in RNAs that are somehow annotated in a similar way. The suite strings and $\eta - \theta$ binning strings for 4,950 structures were calculated in all datasets. In Table 3.1, the distribution of suite codes is shown.

As expected, the helical stem suite variants (1a, 1m, 1L, &a) are predominant. In the two representative datasets, the 1a suites account for up to 60% of all suites, its three satellite clusters contain together another 5%. In SCOR these numbers are very close to that, indicating that the 1a backbone conformation is apt to form many of the motifs annotated there (verified by visual inspection of the primary suite strings). In TRNA the number of 1a is lower (45%). This is a common feature of the tRNA fold, as this observation is the same for all tRNA suite strings. In turn, some of the other suites are more highly represented. In particular, 1L, 1c, 1m, 2g, 4d, 6d, and 1t seem to play an important structural role in tRNA.

The total number of all four kinds of invalid suites ('tt', 'oo','!!', and '−−') are 25.25% in the tRNA set, 12.00% in SCOR, and 14.60%/17.36% in the RNADB05 and HIRES datasets, respectively. At first, the latter seems surprising, because one would expect fewer errors in high resolution structures. The percentage is mainly caused by 3.4% residues with missing atoms. The remaining 13.9% are caused by “triaged” dihedral angles, and by outlier suites

	TRNA	SCOR	RNADB05	HIRES		TRNA	SCOR	RNADB05	HIRES
!!	0.0221	0.0167	0.0100	0.0094	&a	0.0188	0.0252	0.0170	0.0119
−−	0.0005	0.0015	0.0094	0.0343	0a	0.0007	0.0047	0.0041	0.0034
0b	0.0007	0.0012	0.0020	0.0011	1L	0.0422	0.0252	0.0269	0.0201
1[	0.0077	0.0065	0.0078	0.0057	1a	0.4504	0.5760	0.5943	0.6015
1b	0.0110	0.0165	0.0202	0.0244	1c	0.0769	0.0426	0.0477	0.0471
1e	0.0045	0.0063	0.0049	0.0068	1f	0.0098	0.0058	0.0044	0.0023
1g	0.0226	0.0217	0.0127	0.0105	1m	0.0314	0.0177	0.0111	0.0071
1t	0.0046	0.0019	0.0025	0.0031	1z	0.0001	0.0029	0.0023	0.0011
2[	0.0011	0.0057	0.0048	0.0026	2a	0.0117	0.0110	0.0109	0.0122
2g	0.0056	0.0007	0.0015	0.0017	2h	0.0017	0.0010	0.0019	0.0011
2o	0.0003	0.0007	0.0009	0.0020	2u	0.0016	0.0005	0.0009	0.0020
3a	0.0045	0.0084	0.0038	0.0020	3b	0.0004	0.0022	0.0022	0.0014
3d	0.0026	0.0056	0.0027	0.0011	4a	0.0003	0.0026	0.0020	0.0020
4b	0.0023	0.0028	0.0045	0.0048	4d	0.0042	0.0012	0.0017	0.0017
4n	0.0003	0.0013	0.0019	0.0017	4p	0.0017	0.0021	0.0019	0.0014
5d	0.0009	0.0029	0.0019	0.0017	5j	0.0007	0.0020	0.0016	0.0014
5n	0.0012	0.0023	0.0010	0.0009	5p	0.0007	0.0008	0.0011	0.0009
5q	0.0004	0.0006	0.0005	0.0003	6d	0.0057	0.0020	0.0030	0.0045
6g	0.0001	0.0032	0.0033	0.0028	6j	0.0003	0.0008	0.0008	0.0006
6p	0.0003	0.0052	0.0044	0.0043	7a	0.0078	0.0076	0.0043	0.0014
7d	0.0042	0.0046	0.0027	0.0011	7p	0.0020	0.0029	0.0028	0.0034
7r	0.0012	0.0023	0.0017	0.0000	8d	0.0005	0.0026	0.0020	0.0000
9a	0.0019	0.0052	0.0042	0.0051	oo	0.1179	0.0766	0.0723	0.0590
tt	0.1120	0.0352	0.0543	0.0709					

Table 3.1: Ratio of suite codes, as they occur in the four datasets examined here. The table is filled with number of suites of a particular kind, divided by the total number of suites (including outliers) for the corresponding dataset.

for which no suitable cluster could be found. An interpretation of this is that these are unusual backbone conformations which are only visible at a better resolution – in low-resolution structures they probably get smoothed out by the refinement process. In SCOR, the number of invalid suites is much lower. It is clearly biased by the manual selection of motifs, which by definition must occur in well-defined regions.

In the tRNA set, the high error rate was examined in more detail. It appears that the three loop regions contain many conformations that do not fit in any cluster (resulting in 'oo' or 'tt' suites in a row for some structures). This can be a result of strong constraints in the structure during the refinement or by interaction with other molecules. In the high resolution tRNA entry with PDB id 1ehz, the rate of triaged and outlier suites is lower than in the RNADB05 and HIRES sets and the clusters of outliers do not occur here. It is unclear whether modified bases contribute to the problem, but in the examined high-resolution structures this was no problem either. This observation indicates that the lower resolution RNA structures are to be treated with caution.

Analysis of SCOR motifs

The 4,601 motifs from SCOR were divided into a 20% training set and a 80% test set. The training motifs were stored in the suffix tree, and the test motifs searched in it by all their subsequences of 12 characters.

One should assume that e.g. loops of a given type should have similar backbone conformations. Therefore it was investigated into which motifs can be identified this way, and whether they are distinct from other motifs. It was counted how many motifs from the test set could be correctly identified based on matches of their suite strings. In Figure 3.2, the sensitivity and specificity of this analysis is given for each motif class separately.

It turns out, that the predictability of the SCOR motifs is low. While the specificity is above 0.6 for almost all classes examined, and at 1.0 for many of them, the sensitivity covers almost the entire range from zero to one. The reason is a high number of false negatives in each class. To find out where these come from, the suite strings of several classes were inspected in more detail:

The '180 degree turn' class consists of 24 motifs. 17 of them are just two suites (three residues) long, all having the suite string '4b6p'. The remaining 7 contain five suites, which are small variations of '1a3a1g9a1a'. These two groups fully correspond to two homologous positions in different structures of the 23S rRNA (1874-1876 for the first, and 1789-1794 for the second). A similar effect can be observed for many other motifs like '3 non-WC base pair', 'About 90 Degree Turn With All Bases Simply Stacked', and 'Multiple Twist'.

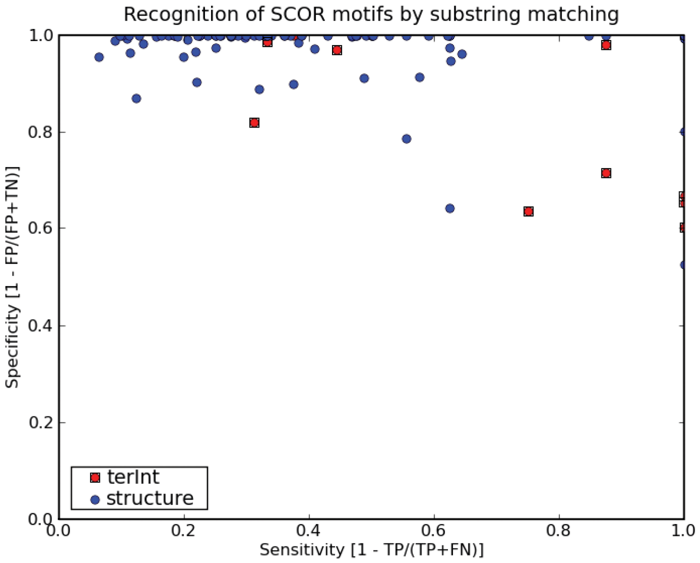


Figure 3.2: SCOR motifs recognized by substring matching. The entire set of SCOR motifs was divided into a 20% training set and an 80% test set. The number of correctly matched substrings of length 12 (or the entire motif, if it was shorter), the number of matches from different SCOR motifs, and the total number of motif pairs compared were used to calculate the sensitivity and specificity of the search.

In other cases, like the 'Ustk stack swap' motif, even more variations can be found.

On the positive side, it has to be noted that the homologous motifs can be recognized well from as few as 2-4 suites, and their structures are conserved. As stated above, the manual selection of motifs probably facilitates this.

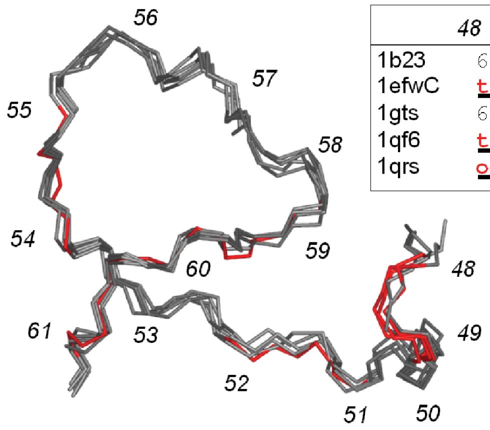
There were no examples found, where two non-homologous motifs belonging to the same class can be identified on the bases of their suites alone. One of the reasons for this observation is that the rules upon which SCOR motifs have been annotated, are based on singular decisions made by experts. It appears, that the base pairing/secondary structure scheme that is specific for a particular motif class, does not impose a constraint on the backbone strong enough to allow a prediction. On the other hand, this implies that in the RNA backbone, an independent set of frequently occurring conformations could exist that has not been described.

Similarities among tRNA

Next, a set of 102 tRNA structures with a well-defined backbone structures was examined. Because all tRNA structures have a highly conserved tertiary structure, one would expect this to be represented in the suite strings as well.

In the TRNA dataset, several suites are over-represented compared to the RNADB05 and HIRES sets (particularly '6d', '2g', '7d', '1f', '1c' and '1L'). These can be found in corresponding positions of most tRNAs. A couple of D-loops from tRNA structures were locally aligned with the corresponding suite strings in Figure 3.3. While each backbone follows the loop along the same path, there are several small differences in the suite codes. These include local variants, often replacing one suite by one close in the 7D dihedral space (e.g. the '1a'–'1L' and '1m'–'1[' exchanges). The structures are also occasionally interrupted by outlier suites. These outliers are visible, but hardly distinguishable in the visualization. They do not alter the direction of the backbone and by no means disrupt the loop structure. Rather, it seems that many of them are results of improper refinement or low structure quality, as high-resolution structures such as PDB-code 1ehz and PDB-code 1b23 are less affected by this. One important conjecture of this is, that the suite codes are a very detailed description of tRNA backbone structure. It is apparently not suitable to describe a well-defined structure such as the D-loop in a general and unambiguous way. For the same loop trace, many combinations of suites are possible.

Another observation is that up to half of the D-loop suites are of the '1a' type, which was described by [Richardson et al., 2008] as the conformer forming 'A-form helices'. The D-loop contains a noncanonical base pair between



	48	49	50	51	52	53	54	55	56	57	58	59	60	61
1b23	6d	1a	1a	1c	1a	1a	1a	1g	1a	1[	4d	1b	2a	
1efwC	tt	1a	1a	**1a**	1a	1a	tt	1g	1a	1[	4d	1b	2a	
1gts	6d	1a	1a	oo	1a	1a	1a	1g	1a	1[	4d	**1t**	2a	
1qf6	tt	1a	1a	1c	1L	1a	1a	1g	1a	1[	4d	1b	2a	
1qrs	oo	1a	1a	1c	1L	1a	1a	1g	**1L**	**1m**	tt	oo	oo	

Figure 3.3: The backbone of the dihydrouridine loops from the tRNA structures with PDB-codes: 1b23, 1efwC, 1gts, 1qf6, and 1qrs superimposed by their backbone atoms. The labels indicate the residue numbers. The suite codes of the dihydrouridine loops are described in the table on the right. Outlier suites are underlined – valid, but singleton suite codes at a given position are highlighted in bold case.

residues 54 and 58, and two adjacent GC base pairs (53-61 and 52-62). But apart from that, many of the bases are involved in tertiary stacking (57, 58) and base pairing (59, 60) interactions. In total, the D-loop stem is more than a simple helix, showing that the abundant 1a suite can accommodate different structural roles.

Although it was not attempted to align all structures explicitly, this seems feasible from these observations, and can be expected to result in a consensus alignment of suites. A more detailed analysis could be used to identify individual conformations of tRNA at a high level of detail.

An all-against-all search of subsequences of all tRNA suite strings was performed using the suffix array, and the n-gram algorithm, as described in Section 3.1.2. In Table 3.2, the numbers of hits found for different word lengths are given.

The tRNA dataset is different enough among itself, that in average only 69 other structures contain a sufficient number of matching n-grams. But, for structures found, the number of words within one hit is high. With increasing word length, the number of hit structures decreases continuously. This is expected as it gets increasingly difficult to find a longer word in the set of suite strings, because each of the occasional variations will disrupt the search for a local match. The number of words found within a structure drops correspond-

	TRNA		RNADB05		HIRES	
n-gram length	number hits	score	number hits	score	number hits	score
4	6,824	5.4	27,978	6.1	10,543	3.7
6	6,732	13.0	22,543	10.6	10,111	7.1
8	6,386	19.1	17,674	14.9	8,917	10.9
10	5,381	24.5	13,657	16.6	6,497	16.5
12	3,812	38.1	10,436	20.4	4,823	20.8
14	2,817	60.9	6,504	30.7	3,321	34.3
16	1,990	96.0	3,554	45.2	2,683	46.8
18	1,542	140.3	2,376	62.0	2,001	59.2
20	1,306	175.4	1,443	86.7	1,283	86.2

Table 3.2: Results of the all-against-all search in the TRNA, RNADB05, and HIRES datasets using the n-gram approach. The column "total hits" indicates how many exactly matching n-grams were found for the given word length. "score" gives the average score for these hits. The score is calculated by the sum of the inverse frequencies from Table 3.1 for the matching n-gram.

ingly at first, but starts to rise again at a word length of 16 (data not shown). This observation can be explained by the fact that these hits are only occurring in a few but highly similar tRNA structures, where little or no variation occurs. It can be concluded that a word size of 12 or 14 is optimal to find similarities within the set with as little background noise as possible, and at the same time not restricting the search to almost-identical structures.

The outcome of the all-against-all search has been visualized in Figure 3.4 (TRNA depicted left). There, the normalized number of word hits for a given pair of structures is plotted. This indicates that an overall level of similarity exists between most pairs of tRNAs. The bright spots result from a group of few highly similar tRNA structures (the ones still remaining with word size 20). The dark regions (the lines at 31, and several ones between 56 and 68) are structures with very low similarity. The structures in this region (among others, PDB-codes: 1yl4, 2ow8, 2v46, 3tra) were examined more closely. It turned out that these contain a much higher proportion (up to 40%) of outlier and erroneous suite codes. Three of the examples are structures of tRNAs bound to ribosomes, having resolutions of 3.7 Å and higher. The fourth (PDB-code: 3tra) is alone, but it also has been determined at an inferior resolution. This clearly shows that the suite nomenclature is of very limited use for non-high-resolution structures.

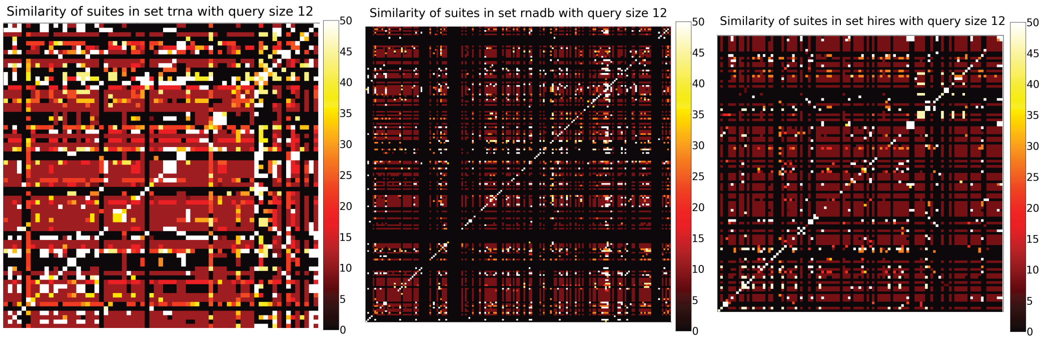


Figure 3.4: Scores of the all-against-all search in the a) TRNA (left), b) RNADB05 (middle), and c) HIRES (right) datasets. On each axis, the structures used are sorted according to their PDB-code. The color indicates the score found for a particular structure-structure-pair. The scaling was chosen such as that dark areas correspond to repeating '1a' matches. The higher the score, the more uncommon suites a particular hit contains. The results shown here are for n-grams of length 12.

Similarities in the representative RNA sets

To assess whether these observations are meaningful, both the 107 high - resolution structures and the 254 structures from the RNADB05 set were compared to each other. The number of hits found is described in Table 3.2. The according similarity maps are depicted in Figure 3.4.

At first, it is observed that some of the suite strings in the datasets were too short to match anything (empty rows/columns and an interrupted diagonal in the heat map). Also, both the HIRES and RNADB05 datasets contained a number of sequences with trivial structures, consisting of '1a'-repeats and not much more. The scoring also depends on the length of the query string and therefore the matrices must not necessarily be symmetric.

In Figure 3.4, it is clearly visible that the overall number of structures in RNADB05 and HIRES with detected similarities drops more sharply compared to the TRNA set. In the same way, the total number of hits changes. Even though the RNADB05 set is larger, only few hit structures remain there at word size 20 (also see Table 3.2). One reason for that is that the average size of both reference datasets is smaller, as they contain many hairpin loops and other short RNA.

In both reference sets, the number of A-form helical stems (repeating regions consisting of '1a' suites) is higher, and they are practically excluded from the evaluation by the scoring function. This leaves only a fraction of hits in the reference compared to the tRNA set. In tRNA not only a higher number of hits exists, but they are also less random because they consist of less frequently occurring suites. This shows that the similarity among tRNAs is non-random, which can be taken as a proof of concept for the method.

One structure in the RNADB05 set – rr0082H09, the 23S subunit of the ribosome – was matched by almost any other from this database. The structural variety in this single structure easily matches that of the remaining dataset taken together, and any motif found somewhere else is probably found there as well (see the white vertical line in Figure 3.4 at dataset RNADB05).

Interestingly, when searching for a set of local RNA structures other than helical stems with either of the methods, non-homologous hits are found. This works for: a) an internal loop of the SRP and the ribosomal SSU, b) a biotin-binding pseudoknot and the tRNA, and c) a tRNA and the E-loop from 5S-RNA.

3.1.4 Discussion

Geometrically, the suite representation does not cover variations that could occur in the bond lengths and flat angles of the RNA backbone. While bond lengths have a very narrow distribution throughout all structure files, bond angles show significant variation. This means that there is a degree of freedom that makes it impossible to rebuild RNA structures from a string, even if the suite nomenclature would determine the dihedrals with perfect precision.

There are two obvious possibilities to resolve this:

1. Encode the flat angles in a similar way as the suites.
2. Encode base-base interactions in the string in order to constrain the structure, and use a 3D modeling procedure subsequently.

The second method is assumed to be more promising, because it would include those interactions that shape the function of RNA instead of restricting the structure of RNA to the backbone alone. Such a reconstruction of structures from a descriptive grammar (not string-based) was demonstrated already in [Parisien and Major, 2008]. Another implication of this approach would be, that if an RNA has in some region no further constraints, it may be structurally flexible. Therefore, the second approach would indirectly encode the flexibility.

Having a rapid method for string-based motif recognition has a number of potential applications. First, it could be used to systematically find frequently occurring backbone motifs in RNA structures – as it has been demonstrated here. Further, it can be used to sample big numbers of backbone conformations in order to generate native-like RNA backbones which could be modeled subsequently. Finally, it allows on-the-fly evaluation of RNA models which are generated during manual structure modeling or automatic refinement. The combination of this technique with more elaborate string representations would impose further improvement. Therefore it can be assumed that it is possible to accurately re-model the structure of RNA from a string representation by including additional structural features like base pairs, base stacking, or even tertiary interactions with energy minimization instead of extensive probing of the local conformational space.

The $\eta - \theta$ binning approach was shown to produce too many different local conformations for an effective substring matching. One could argue that by decreasing the number of bins, the matching could be improved. But, it has been shown earlier, that the pseudo torsion angles contain specific regions that are characteristic for some structural motifs [Duarte and Pyle, 1998]. Decreasing the bin size would ignore these and therefore be hopelessly inaccurate. Therefore, either explicit clusters in the pseudo torsion space would have to be defined or string matching techniques allowing for more inexact matches than the current suffix array would be necessary. A fuzzier search method could improve the usefulness of the suite codes as well. In particular, this could eliminate the adversary effects of the occasionally occurring erroneous or undefined suites. Practically, this could be implemented as a classical similarity matrix between the suite codes, and for the beginning, its values could simply be based on a normalized 7D distance between the 46 suite clusters. Given the performance of the suffix array the analysis presented here could easily be extended to the entire NDB [Berman et al., 2002]. Identifying structures that should be expected to be similar (e.g. based on their function) is more challenging, if one does not want to rely on sequence similarity alone.

3.1.5 Conclusions

The first approach that uses an indexing technique to scan the structural space of RNA was presented. The indexing was implemented using suite codes and an $\eta - \theta$ binning approach and tested on four distinct datasets. It could be shown that this approach can be used to rapidly identify similar substructures. This has applications not only for querying the RNA space but also for the modeling of RNAs by rapidly predicting possible conformations and in turn on-the-fly evaluation of proposed RNA models regarding structural and functional similarities. All datasets as well as the source code is freely available from http://suiterna.sourceforge.net.

3.2 Macromolecular similarity screening

This part presents a generalized approach for the fast structural alignment of thousands of macromolecular structures. The method uses string representations of a macromolecular structure and a hash table that stores n-grams of a certain size for searching. To this end, macromolecular structure-to-string translators were implemented for protein and RNA structures. A query against the index is performed in two hierarchical steps to unite speed and precision. In the first step the query structure is translated into n-grams, and all target structures containing these n-grams are retrieved from the hash table. In the second step all corresponding n-grams of the query and each target structure are subsequently aligned, and after each alignment a score is calculated based on the matching n-grams of query and target. The extendable framework enables the user to query and structurally align thousands of protein and RNA structures on a commodity machine and is available as open source from http://lajolla.sf.net.

3.2.1 Introduction

Macromolecules and their function

The function of macromolecules is determined by their three-dimensional (3D) structure. This 3D structure allows for a specific binding of small compounds like drugs, metabolites, or other macromolecules such as RNA and proteins. This binding process is crucial for cell signaling and of great interest for understanding the cellular apparatus and the development of new treatments for diseases. Determining the structure of a macromolecule (protein, RNA) and thus the coordinates of the residues in atomic detail was and still is a significant procedure. The first structures, hemoglobin and myoglobin, were determined 1958 by Kendrew et al. [Kendrew et al., 1958]. Since then progress has been made towards a faster determination of macromolecular structures. However, for many macromolecules it is still impossible to determine the complete structure [Scheerer et al., 2008]. The principal repository for the coordinates of macromolecular structures is the wwPDB archive [Berman et al., 2007]. As of November 2008 the wwPDB stores well over 50,000 structures consisting of roughly 1,500 RNA structures including protein - RNA complexes and 48,000 proteins. In recent years various structural genomics initiatives were started that aimed towards a fast, high-density determination of thousands of macro-molecular structures [Service, 2008, Levitt, 2007]. These initiatives led to around 1500 structures with unknown functions. The annotation of macromolecules can be carried out on different levels, however, the manual annotation of those structures is often not feasible despite best efforts [Rother et al., 2005, Andreeva et al., 2008]. The fastest way to determine the function is to use the sequence of its building blocks (the primary structure) alone, and search this sequence against a database of annotated structures where the function can be subsequently inferred. This approach generally works well when the sequences are highly similar but sometimes fails [He et al., 2008]. A more accurate way to annotate is to use 3D information. Many methods try to identify secondary structure elements and align them with each other. These approaches are often sequence-independent and therefore not subject to failure because of relative sequence similarity [Cheek et al., 2004, Shindyalov and Bourne, 1998]. A general fact for both protein and RNA structural alignment is that there often cannot be a single best solution to align two or more structures. The best solution is always the best given a certain man-made optimization criteria, nicely explained by [Sippl and Wiederstein, 2008].

Protein function and similarity

The importance of structural alignments of protein structures is based on the fact that structural motifs (folds) contained in the structure reveal important biochemical functions [Andreeva et al., 2008]. For instance the so-called "Rossman fold" is a strong indication for the binding of nucleotide derivatives [Rao and Rossmann, 1973]. For performance reasons, many computational algorithms work on the sequence level, while also taking into account the 3D secondary structure as guidance [Shindyalov and Bourne, 1998, Zhang and Skolnick, 2005]. In many scenarios this approach proves to be fast and accurate enough. However, given the already mentioned fact that a similar sequence does not necessarily mean a structural similarity there are a growing number of approaches that use pure 3D information to overcome this disadvantage [Guerler and Knapp, 2008, Ilyin et al., 2004]. In this regard the authors want to especially stress the SSM project, which is the first software fast enough to search the whole PDB within minutes with a high accuracy based on an abstraction of the 3D structure [Krissinel and Henrick, 2004]. Wikipedia currently lists more than 50 different approaches for protein alignment (http://en.wikipedia.org/wiki/Structural_alignment_software). A detailed comparison of algorithms and approaches in the field is presented by [Kolodny et al., 2005, Novotny et al., 2004]. The approach presented can be adjusted regarding speed and precision / coverage. A schema frequently used to express the backbone of a protein or RNA is to use torsion angles between a well-defined set of atoms. The torsion angles between consecutive amino acids became famous when Ramachandran et al. published the analysis of the ϕ (phi) and ψ (psi) torsion angles (Definition 3.2.2) of protein chains in

1963 [Ramachandran et al., 1963]. Ramachandran showed that the usage of ϕ and ψ angles allows for a clear separation of secondary structure elements (Figure 3.5). This in turn allows us to judge whether amino acids belong to a certain class of secondary structure elements like α-helices or β-sheets. This notion was frequently applied in the abstraction and search of similar protein structures and is often used together with techniques such as suffix trees and suffix arrays (among others [Guyon et al., 2004, Täubig et al., 2006, Friedberg et al., 2007, Lo et al., 2007, Gao and Zaki, 2008]). An interesting approach in the field of protein-protein interaction is proposed by Günther et al. where known motifs of interacting domains are used to predict potential interactions of novel proteins [Günther et al., 2007].

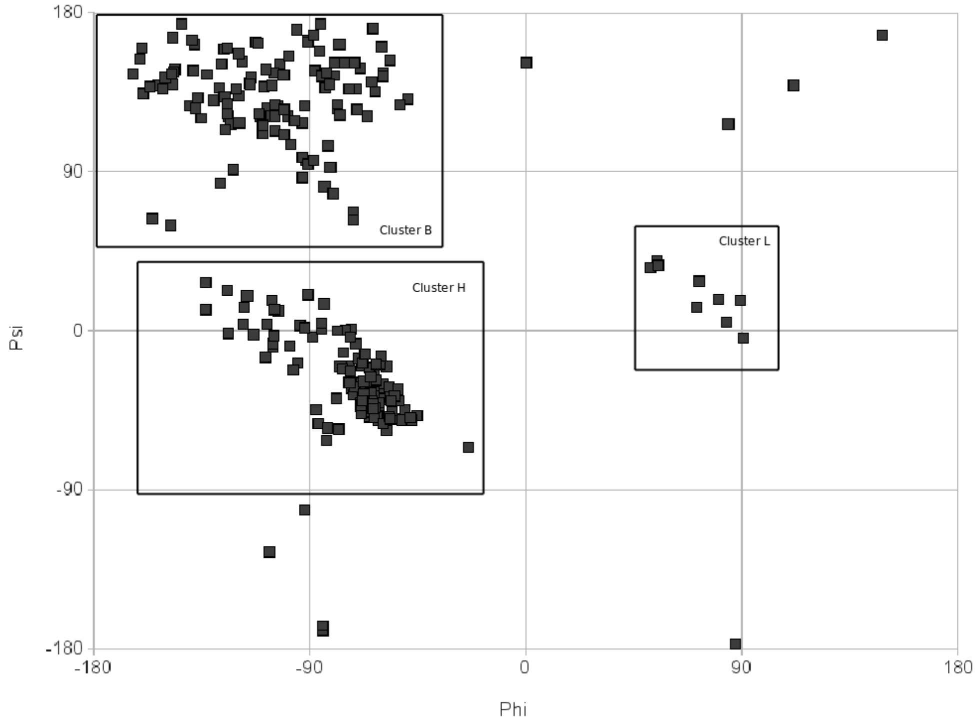


Figure 3.5: The Ramachandran ($\phi - \psi$ torsion angle) plot of a Thymidylate Synthase (PDB-ID: 1AXW). The cluster in the upper left corresponds to β-sheets, the cluster in the middle left corresponds to α-helices, and the small cluster in the middle right represents left handed helices. The main clusters (B,H,L) are used to translate a protein structure into a string.

RNA function and similarity

In recent years, RNA gained attention due to the discovery of their heavy involvement in the regulatory apparatus of the cell [Laederach, 2007]. Apart from the fact that a relatively small amount of RNA structures are contained in the wwPDB they are nevertheless of growing importance [Tamura et al., 2004, Abraham et al., 2008]. As this field is relatively young, there are only a few structural RNA alignment methods available [Guyon et al., 2004, Chang et al., 2008, Capriotti and Marti-Renom, 2008], but interest in the structure of RNA is rapidly growing. It has to be noted that there is currently no methodology available that allows for the querying of an RNA motif against all RNA structures in real time, as it is provided by SSM for the world of proteins. A schema to express the backbone of an RNA is the usage of η (eta) and θ (theta) pseudo torsion angles (Definition 3.2.1) - representing each nucleotide in a chain by two angles. In a thorough analysis of this pseudo torsion representation, eight main classes of conformations have been identified, and this information could be exploited to highlight important features of the RNA structures [Wadley et al., 2007]. A more detailed approach was taken by Richardson et al. where a set of 46 nucleotide conformations is determined based on the seven torsion angles present in a ribose-to-ribose (suite) unit [Richardson et al., 2008]. This representation implicitly includes the pucker of the ribose, and it is detailed enough to track down the conformations in local motifs such as GNRA tetraloops. To do this, however, it is necessary to have well-resolved RNA structures available. Both of these high level abstractions are limited to the RNA backbone, and their accuracy is not sufficient to reconstruct an RNA structure from the string representation alone. Nevertheless, adding information such as canonical and noncanonical base pairs, as well as base stacking provides sufficient input to assemble RNA tertiary structures from such a combined descriptor alone. Recent progress in the field of RNA structure prediction demonstrates the feasibility of this approach [Parisien and Major, 2008], except that the attempt to write the structural descriptor as a string has not been made. The RNA Ontology Consortium is currently standardizing the component descriptors of RNA chains in order to facilitate further work on the subject [Leontis et al., 2006].

Scope

The aim of this work is to propose a novel approach for the fast hierarchical search of similarities in thousands of macromolecular structures. The method is based on a fast index structure, derived from the field of classical string alignment [Gusfield, 1997]. But unlike classical sequence-based search methods, the strings can represent structural features of the 3D structure and are sequence independent. Thus, this approach has the potential to be as fast as sequence-based approaches with the precision of structural alignment methods. The authors want to stress the term “fast,” as many approaches currently work in a one-against-one mode. The proposed method, materialized in the framework LaJolla, is easily extendable with new chain-to-string translators. The aim of this publication is to present this approach and the software package as a potentially useful tool for many domains. The in-depth validation for individual domains such as protein and RNA similarity, protein-protein interaction and protein-small compound docking is subject to publications in journals of the corresponding communities.

3.2.2 Material and methods

In a nutshell

The proposed approach performs a search for local structural motifs in a set of 3D structures of macromolecules. The basic ideas for this approach originate in [Bauer et al., 2008], where different possibilities to represent RNA structures as a reduced alphabet and the possibility of storing and querying that alphabet in suffix-based indices were analyzed. The paper clearly shows that the proposed methodology of suite codes [Richardson et al., 2008] is too narrow for an RNA search. To tackle this drawback the notion of n-grams was used. However, n-grams can be used more easily in hash tables than in the originally implemented suffix-based structures. It was also if interest in how the sophisticated suite methodology compares to a simple $\eta - \theta$ torsion angle discretization. A novel development that enhances the practical application is that LaJolla performs a final 3D superposition to remove statistical artifacts that have no real 3D significance. During the development, it turned out that the approach is not only useful for RNA structures but also for proteins using the respective transformers (for instance $\phi - \psi$).

String representation of linear polymers

This approach is based on the simple observation that macromolecular structures share a common property: They are made up of chains formed by molecular building blocks, and possess a linear molecular backbone with repeating units. This property allows for the application of abstractions that are able to translate these macromolecules into a one-dimensional (1D) linear representation (Figure 3.6). This in turn allows for the use of efficient algorithms deriving from the field of string matching and text mining [Gusfield, 1997].

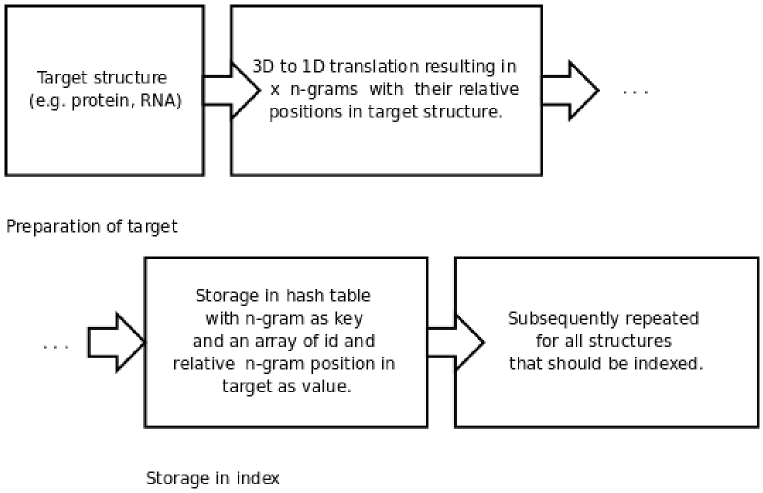


Figure 3.6: An informal diagram showing the steps issued for the initial generation of the hash table used as index structure for the structural alignment.

From torsion angles to a string

A simple translation procedure from a 3D structure to a 1D string is to use the sequence of nucleic acids or amino acids. However, as already stated in the introduction, a high similarity in sequence does not imply that the structures are similar. To overcome this, the default procedure uses torsion angles between defined atoms. In the case of RNA structures, the translator maps the residues to $\eta - \theta$ pseudo torsion angles (Definition 3.2.1). In case of proteins $\phi - \psi$ torsion angles are used (Definition 3.2.2).

Definition 3.2.1 *Given three consecutive nucleotides N1, N2, N3 of a nucleic acid chain. Let η be the torsion angle defined by atoms ($N1_{C4'}$, $N2_P$) and ($N2_{C4'}$, $N3_P$). Let θ be the torsion angle defined by atoms ($N2_P$, $N2_{C4'}$) and ($N3_P$, $N3_{C4'}$).*

Definition 3.2.2 *Given three consecutive amino acids A1, A2, A3 of a polypeptide chain let ϕ be the torsion angle defined by atoms ($A1_C$,$A2_N$) and ($A2_{CA}$, $A2_C$). Let ψ be the torsion angle defined by atoms ($A2_N$,$A2_{CA}$) and ($A2_C$, $A3_N$).*

Once a sequence of torsion angles is generated it can be translated into a sequence of characters using any function. In the case of proteins as well as for RNAs the main clusters of the dihedral angle plots are translated into distinct characters (see also Figure 3.5).

The result of that translation step is a string where single characters represent the torsion angles of the chain residues and therefore the macromolecule as a whole. Traditional string matching algorithms can subsequently be applied, enabling the user to index and to search for macromolecular structures.

An n-gram based index structure for fast searches

A hash table is a data structure that stores key - value pairs. A value can be a character, a string or an arbitrary object. The key is generated by a mathematical function (hashing function) that translates the value into the key. This key in turn allows us to retrieve the value from a hash table in an average run time of O(1) [Dietzfelbinger et al., 1988]. There are two characteristics of hash tables that have major influence on the run time. First, not all hashing functions necessarily yield unique results, subsequently, collisions have to be resolved by chaining values or by other approaches. Second, to obtain the average run time of O(1) an average load factor has to be kept, and a so-called rehashing has to be issued if the load factor goes below a certain threshold. A good general introduction to the field is given by [Cormen et al., 2003]. To conclude, a hash table allows for a fast determination if certain strings are contained in the index. However, storing the complete sequence of a chain (e.g. discrete $\eta - \theta$ values) as value in the hash table does not make much sense because it would only allow searches for exact matches of whole structures that virtually never occur. To overcome this disadvantage it is useful to store so called n-grams (also: q-grams) of a sequence in the hash table [Burkhardt et al., 1999]. An n-gram is a string of length n. All n-grams of a string m are all sub-strings of length n of m. For example all 2-grams of the string ALICE are AL, LI, IC and CE. N-grams are widely used as a statistical tool to define the relatedness of two strings. Google's "Did you mean: ..." feature is a classic example of that. But n-grams can also be used as method for fuzzy string alignment. If the string ALICE is searched in the string ALITE using 2-grams, then two 2-grams are found, two missed, and an alignment can be proposed by this approach.

Generating and searching the index

For searching, a hash table is generated from all n-grams of all target structures, in which n-grams generated from the query structure are searched. Generating the n-gram based index is a straightforward process (Figure 3.6). All target structures (chains) have to be translated subsequently to strings using a structure-to-string translator. The n-grams of each target structure are stored in the hash table. It has to be noted that the positions of the n-grams of query and target are stored as well, making it feasible to perform a 3D alignment for scoring and refinement. Searching a structure (query) in the index involves the transformation of the query chain into a string and the computation of each n-gram (Figure 3.7). The search results in a certain amount of target structures that have n-grams in common with the query. As these results may be statistical artifacts, a second hierarchical refinement step is applied. In this refinement step, the corresponding n-grams are subsequently aligned and thus anchors of query and target are determined. With that allocation, a superposition of query and target is performed [Kabsch, 1976]. The scoring is carried out by calculating the RMSD (Definition 2.2.1) and a qualitative score, TM-score, as defined in [Zhang and Skolnick, 2007] (Definition 3.2.3). The RMSD alone is not suitable as it does not allow conclusions to be drawn about the number of residues that have been aligned successfully.

Definition 3.2.3

$$\mathrm{TM-score} = \frac{1}{L_{Target}} \sum_{i=1}^{L_{aligned}} \frac{1}{1 + \left(\frac{d_i}{1.24\sqrt[3]{L_{Target}-15}-1.8} \right)^2} \tag{3.2}$$

where L_{Target} and $L_{aligned}$ are the lengths of the target and aligned structure respectively. d_i is the Euclidian distance between the ith pair of residues.

An advantage of the presented approach is that the strings (n-grams) that are being indexed and in turn searched using the hash table can be generated by an arbitrary approach. From the perspective of software engineering it is easily possible to exchange the discussed approach of protein $\phi - \psi$ torsion with the Protein Blocks Method [Tyagi et al., 2008] mentioned in the introduction. For RNA structures it would be easily possible to replace the $\eta - \theta$ torsion angles approach with the notion of suite codes proposed by Richardson et al. [Richardson et al., 2008], or any other representation.

The principal parameters that have an impact on performance and accuracy are the size (n) of the indexed n-grams and complexity of the string a structure to string translator produces. In an extreme case a structure to string translator would produce always the same letter for each angle combination meaning each n-gram of the query will be compared to the each n-gram of the target. A clever translator reduces this by only comparing beta-sheets and helices or even combinations using a longer n-gram size reducing the search time dramatically.

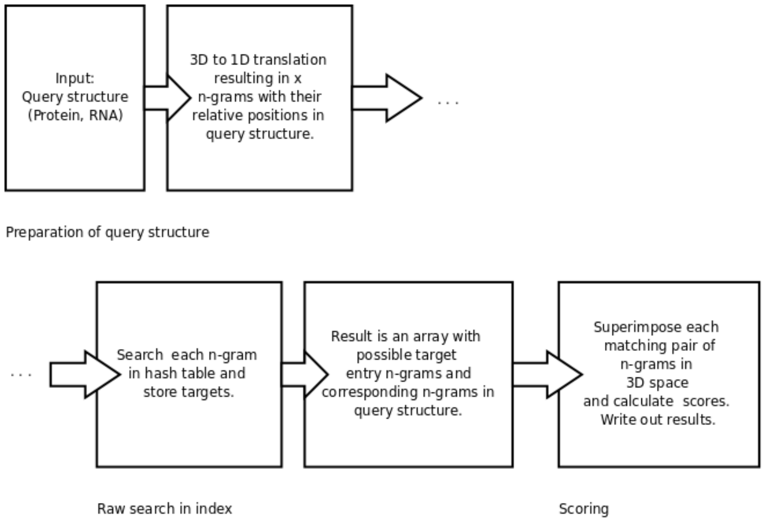


Figure 3.7: An informal diagram showing the steps performed when a query structure is searched against a set of target structures in an index.

Datasets used

The datasets can be downloaded from the project homepage at http://lajolla.sf.net.

tRNA dataset For the analysis of the RNA alignment capabilities of LaJolla, all molecular structures containing a tRNA were retrieved from the NDB database. The dataset was filtered manually, to identify the polymer chains, to identify the functional state of the molecules, and exclude structural fragments. The resulting dataset contains 101 nucleic acid chains, all of which have been resolved by X-Ray crystallography.

Protein benchmarking Two datasets for benchmarking LaJolla in the domain of proteins were used. The first Dataset termed CATH_1258 is derived from the CATH [Cuff et al., 2009]. The S35 subset of CATH version 3.2.0 (CathDomainPdb.S35) was used. From this subset the first entry of each structure at “H” level was chosen. Entries that are singletons regarding their parent topology level were subsequently removed resulting in 1,258 entries. Thus, each of these 1258 structures is classified by CATH and has at least one entry that is classified in the same class - architecture - topology combination. For the evaluation of the performance 100 structures were randomly picked from the S35 subset of CATH termed CATH_100.

3.2.3 Results

The following results were performed with the default settings of LaJolla version 2.0. Results below a TM-score of 0.2 are neglected. The standard translators were used, for proteins *BetterOptimizedPhiPsiTranslator*, for RNA structures *OptimizedStructureToEtaThetaCharacterTransformer*. For comparison to the state of the art CE (version 2004/10/07) [Shindyalov and Bourne, 1998] and TM-align (64bit version 2005/06/01) [Zhang and Skolnick, 2005] were chosen.

Performance

As the CATH_1258 dataset was executed on a distributed environment it is not possible to take these runtimes. To this end, the CATH_100 dataset was used and executed in an all against all manner for LaJolla (n-gram sizes 10, 15, 20, 25, 30), CE and TM-align (Figure 3.8). The tests were executed on standard hardware with an Operton 2.2 GHz with only one CPU enabled. The histogram points out that LaJolla is fast when using larger n-gram sizes. CE is the slowest method.

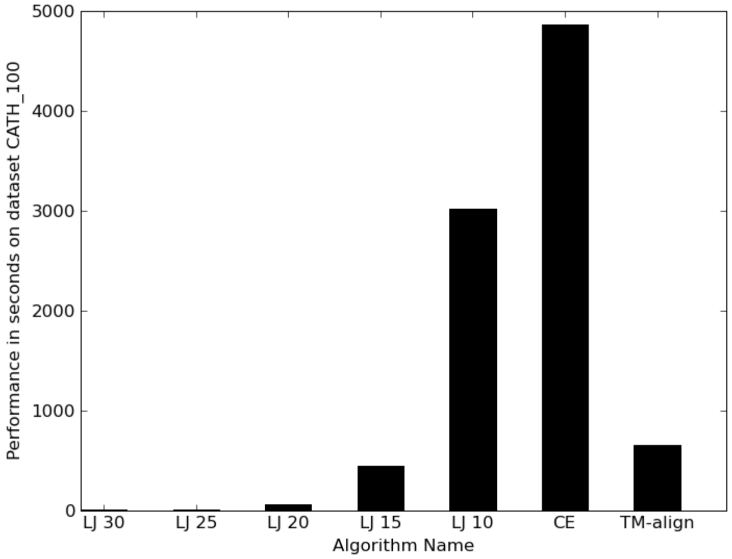


Figure 3.8: Runtimes of the algorithms LaJolla (LJ) (n-gram size 10,15,20,25,30), CE and TM-align when performing an all against all comparison of 100 structures (dataset CATH_100).

RNA retrieval

A multiple structural alignment of 101 tRNA chains was performed using the $\eta - \theta$ angle representation implemented in LaJolla. The tRNA molecule was chosen for this task, as it possesses a highly conserved tertiary structure that is straightforward to recognize and to validate. Despite that, it contains many local structural variations, and changes its conformation depending on its functional state (see [Giegé, 2008] for a review). Finally, a high number of structures of different quality are available for this family of RNA. The all-against-all search in LaJolla resulted in $101^2 = 10201$ queries that were performed with an n-gram size of 10. The run resulted in 10195 local alignments returned by the program. To validate the results, it was checked how well the query and target structures are superimposed by the method. By manual inspection it was determined that finding at least 30 residues close to each other, or a TM-score higher than 0.25, were sufficient criteria to tell apart correct global superpositions and mere local similarities. Using these criteria, 9,237 (90.5%) superpositions were done successfully. The full list of examined RNA chains and average RMSD, TM-score, and number of aligned residues are given in Tables 3.3 and 3.4. Inspecting the results in detail, it was found that for the RNA chains 2nre/F and 1j2b/C+D more than 60% of the superpositions failed. In both cases, the RNAs are forced by a base modifying enzyme into an unusual conformation (pseudouridine synthase and archaeosine transglycosylase, respectively). In the case of 1j2b, an entire arm of the tRNA changes its conformation (called lambda-form tRNA). Other functional states of the matched tRNA molecules shows little influence on the number of hits. By far the most abundant state available is tRNA bound to aaRS proteins (56 chains), and it has on average 91.8 correct hits found by LaJolla. The next most frequent group are ribosomes (26 chains), with 96.3 correct hits in average. In total, there are on average 91.5 correct hits per chain. The number of alignments found may result from similarities of the functional states, but it was not possible to confirm this as significant - for this, one would expect e.g. tRNA in complex with aaRS to prefer each other in the hit list, and ribosome complexes among each other etc. This was, however, not observed. A bad resolution seems to rather improve the alignability of a structure, as observed on the ribosomes. A simple explanation for this is that the tRNA in many of the ribosomal structures has been constructed by molecular recognition techniques using a standard template - and intricate local variations not detectable in the structures. As a result, it can be stated that, using the default $\eta - \theta$ translator, it is possible to align badly-resolved structures correctly - a feature not attainable by the suite code translator. On the positive side, inspection of the local alignments showed that they are not altogether local. LaJolla finds a series of matching n-grams throughout a pair of structures. Thus, the structural alignments are not based on a local similar substructure common to both molecules, but rather a consensus of many small similarities that add together to the final alignment. Only in the incorrect hits was the alignment confined to some part of the structure.

For analyzing the sensitivity of the tRNA structural alignment, 60 RNA structures annotated in the SCOR database, including 13 tRNAs (taken from [Capriotti and Marti-Renom, 2008]) were compared. It was calculated, how many times the highest scoring structure retrieved by LaJolla has the same class in the 'functional annotation' category. For tRNA, this was the case for 100% of the entries regardless of n-gram size. This shows that tRNA structures being that similar to each other that even a moderately accurate superposition it sufficient to distinguish them from other types of RNA. When considering the accuracy of other functional classes, the retrieval gets much less accurate, with only 53% correctly assigned functional categories (when considering the best of the top five TM-scores, this number rises to 69%). One of the reasons for the observed wrong assignments is that part of the 47 non-tRNA structures express considerable structural variety despite their small size. This sensitivity can be improved by applying a TM-score cutoff, but this may lead to misleading figures because then the tRNA structures will be heavily overrepresented in the data. This points to limits of the SCOR dataset, and suggests that a manual functional annotation of those parts of the PDB not covered by SCOR would be helpful.

Protein retrieval

To evaluate the capabilities of LaJolla in the field of protein retrieval the CATH protein classification was used as standard of truth and compared the results to two other popular algorithms in the field: CE [Shindyalov and Bourne, 1998] and TM-align [Zhang and Skolnick, 2005]. CATH allows us to validate if the results produced by a method are "true" in terms of a similar classification. To this end, the topology level of CATH was chosen. The reduced dataset CATH_1258 ensures that there is at least one other protein on the same topology level. The graphs in Figure 3.9 show how well the classification works in regard to the coverage at a certain scoring cutoff. To assess correct hits, it was counted if the result with the best score was true (TOP 1) and also if a correct result was among the ten best hits (TOP 10). The results show that CE, despite its age, still is a very good method with good overall results. It ranks best when it comes to TOP 1 hits and second when it comes to TOP 10 hits. If one takes into account when the coverage line crosses the percentage of correct TOP1 and TOP 10 hits CE also ranks first. TM-align is

PDB-ID chain	resol.	type	complex with	hits	RMSD	TM-score	% aligned
1b23 R	2.60	tRNA_Cys	Ef-Tu	87	1.75	0.38	36.24
1c0a B	2.40	tRNA_Asp	AspRS	98	1.71	0.48	44.85
1efw C	3.00	tRNA_Asp	AspRS	92	1.76	0.48	44.95
1efw D	3.00	tRNA_Asp	AspRS	95	1.73	0.48	44.29
1ehz A	1.93	tRNA_Phe	uncomplexed	98	1.70	0.52	49.18
1eiy C	3.30	tRNA_Phe	PheRS	65	1.89	0.33	33.23
1euq B	3.10	tRNA_Gln	GlnRS	98	1.71	0.52	46.91
1euy B	2.60	tRNA_Gln	GlnRS	98	1.67	0.52	47.00
1exd B	2.70	tRNA_Gln	GlnRS	99	1.75	0.51	47.07
1f7u B	2.20	tRNA_Arg	ArgRS	98	1.75	0.43	40.52
1f7v B	2.90	tRNA_Arg	ArgRS	98	1.74	0.44	40.66
1ffy T	2.20	tRNA_Ile	IleRS	96	1.69	0.48	44.37
1g59 B	2.40	tRNA_Glu	GluRS	89	1.62	0.49	44.88
1g59 D	2.40	tRNA_Glu	GluRS	88	1.65	0.49	44.54
1gts B	2.80	tRNA_Gln	GlnRS	95	1.70	0.48	44.53
1h3e B	2.90	tRNA_Tyr	TyrRS	97	1.77	0.45	42.83
1h4s T	2.85	tRNA_Pro	ProRS	91	1.68	0.45	38.62
1il2 C	2.60	tRNA_Asp	AspRS	90	1.84	0.45	44.02
1il2 D	2.60	tRNA_Asp	AspRS	96	1.72	0.47	42.05
1j1u B	1.95	tRNA_Tyr	TyrRS	99	1.63	0.50	45.30
1j2b C	3.30	tRNA_Val	archaeosine transglycosylase	38	1.79	0.34	32.84
1j2b D	3.30	tRNA_Val	archaeosine transglycosylase	31	1.80	0.35	32.43
1n77 C	2.40	tRNA_Glu	GluRS	94	1.62	0.49	44.59
1n77 D	2.40	tRNA_Glu	GluRS	91	1.68	0.50	46.08
1n78 C	2.10	tRNA_Glu	GluRS	93	1.62	0.50	45.50
1n78 D	2.10	tRNA_Glu	GluRS	91	1.69	0.50	46.42
1ob2 B	3.35	tRNA_Phe	Ef-Tu	97	1.85	0.43	42.31
1pns V	8.70	tRNA_Phe	70S ribosome	98	1.71	0.53	49.57
1pns W	8.70	tRNA_Phe	70S ribosome	99	1.70	0.50	46.53
1qf6 B	2.90	tRNA_Thr	ThrRS	96	1.71	0.45	41.68
1qrs B	2.60	tRNA_Gln	GlnRS	94	1.69	0.49	45.49
1qrt B	2.70	tRNA_Gln	GlnRS	94	1.70	0.48	44.62
1qru B	3.00	tRNA_Gln	GlnRS	94	1.69	0.49	44.97
1qtq B	2.25	tRNA_Gln	GlnRS	98	1.68	0.49	44.82
1qu2 T	2.20	tRNA_Ile	IleRS	96	1.69	0.48	44.37
1qu3 T	2.90	tRNA_Ile	IleRS	98	1.68	0.49	44.93
1wz2 C	3.21	tRNA_Leu	LeuRS	97	1.75	0.43	40.84
1wz2 D	3.21	tRNA_Leu	LeuRS	97	1.74	0.46	43.21
1yl4 B	5.50	tRNA_Phe	70S ribosome	98	1.83	0.50	48.26
1yl4 C	5.50	tRNA_Phe	70S ribosome	99	1.75	0.50	47.07
1zjw B	2.50	tRNA_Glu	GluRS	98	1.68	0.50	45.61
2ake B	3.10	tRNA_Trp	TrpRS	96	1.67	0.44	40.14
2azx C	2.80	tRNA_Trp	TrpRS	100	1.72	0.50	45.71
2azx D	2.80	tRNA_Trp	TrpRS	100	1.73	0.48	44.01
2b64 V	5.90	tRNA_Phe	70S ribosome	98	1.76	0.47	45.16
2b64 W	5.90	tRNA_Phe	70S ribosome	98	1.82	0.52	49.75
2b9m V	6.76	tRNA_Phe	70S ribosome	98	1.77	0.47	44.88
2b9m W	6.76	tRNA_Phe	70S ribosome	99	1.83	0.48	46.98
2b9o V	6.46	tRNA_Phe	70S ribosome	100	1.78	0.46	44.43
2b9o W	6.46	tRNA_Phe	70S ribosome	98	1.79	0.51	49.07
2bte B	2.90	tRNA_Leu	LeuRS	86	1.88	0.42	40.93
2bte E	2.90	tRNA_Leu	LeuRS	81	1.84	0.41	40.17

Table 3.3: tRNA search result table (part 1).

PDB-ID chain	resol.	type	complex with	hits	RMSD	TM-score	% aligned
2byt B	3.30	tRNA_Leu	LeuRS	72	1.85	0.41	40.15
2byt E	3.30	tRNA_Leu	LeuRS	71	1.85	0.42	40.36
2csx C	2.70	tRNA_Met	MetRS	95	1.68	0.47	44.03
2csx D	2.70	tRNA_Met	MetRS	95	1.66	0.47	43.39
2ct8 C	2.70	tRNA_Met	MetRS	99	1.69	0.47	43.53
2ct8 D	2.70	tRNA_Met	MetRS	97	1.70	0.43	40.05
2cv0 C	2.40	tRNA_Glu	GluRS	93	1.62	0.49	44.53
2cv1 C	2.41	tRNA_Glu	GluRS	93	1.64	0.50	45.99
2cv1 D	2.41	tRNA_Glu	GluRS	91	1.70	0.50	46.78
2cv2 C	2.69	tRNA_Glu	GluRS	92	1.65	0.51	46.75
2cv2 D	2.69	tRNA_Glu	GluRS	91	1.69	0.50	46.31
2d6f E	3.15	tRNA_Gln	GluRS	97	1.80	0.43	40.94
2d6f F	3.15	tRNA_Gln	GluRS	98	1.87	0.41	40.11
2der C	3.10	tRNA_Glu	mnma thiolase	98	1.74	0.48	44.98
2der D	3.10	tRNA_Glu	mnma thiolase	96	1.70	0.50	44.92
2det C	3.40	tRNA_Glu	mnm5s2U-methyltransferase	94	1.72	0.45	40.28
2deu C	3.40	tRNA_Glu	mnm5s2U-methyltransferase	90	1.73	0.43	40.93
2deu D	3.40	tRNA_Glu	mnm5s2U-methyltransferase	89	1.73	0.44	41.02
2dr2 B	3.00	tRNA_Trp	TrpRS	100	1.68	0.43	39.79
2du3 D	2.60	tRNA_Cys	o-phosphoserylRS	95	1.76	0.45	41.51
2du4 C	2.80	tRNA_Cys	o-phosphoserylRS	95	1.78	0.46	42.32
2du5 D	3.20	tRNA_opal	o-phosphoserylRS	93	1.88	0.41	39.22
2du6 D	3.30	tRNA_Amber	o-phosphoserylRS	96	1.89	0.40	38.27
2dxi C	2.20	tRNA_Glu	GluRS	92	1.62	0.49	44.98
2dxi D	2.20	tRNA_Glu	GluRS	88	1.63	0.49	44.78
2fk6 R	2.90	tRNA_Thr	RNase Z	85	1.57	0.52	36.29
2hgi C	5.00	tRNA_fMet	70S ribosome	99	1.69	0.52	48.56
2hgi D	5.00	tRNA_Phe	70S ribosome	86	1.90	0.42	41.56
2hgp B	5.50	tRNA_Phe	70S ribosome	90	1.91	0.44	43.47
2hgp C	5.50	tRNA_Phe	70S ribosome	98	1.77	0.49	46.43
2hgp D	5.50	tRNA_Phe	70S ribosome	90	1.84	0.42	41.07
2hgr C	4.51	tRNA_fMet	70S ribosome	100	1.68	0.51	47.38
2hgr D	4.51	tRNA_Phe	70S ribosome	93	1.89	0.43	42.78
2iy5 T	3.10	tRNA_Phe	PheRS	51	1.95	0.33	33.58
2j00 W	2.80	tRNA_Phe	70S ribosome	97	1.76	0.44	42.02
2j02 V	2.80	tRNA_fMet	70S ribosome	98	1.69	0.49	46.10
2j02 W	2.80	tRNA_Phe	70S ribosome	97	1.80	0.46	44.09
2nre F	4.00	tRNA_Leu	pseudouridine synthase	32	1.56	0.46	33.68
2ow8 0	3.71	tRNA_Phe	70S ribosome	93	1.89	0.42	41.68
2ow8 z	3.71	tRNA_Phe	70S ribosome	90	1.82	0.45	43.30
2qnh 2	3.83	tRNA_Phe	70S ribosome	93	1.84	0.43	41.65
2qnh z	3.83	tRNA_fMet	70S ribosome	100	1.74	0.51	48.56
2tra A	3.00	tRNA_Asp	uncomplexed	98	1.72	0.44	40.72
2v0g B	3.50	tRNA_Leu	LeuRS	69	1.86	0.42	40.74
2v0g F	3.50	tRNA_Leu	LeuRS	69	1.85	0.42	40.22
2v46 W	3.80	tRNA_fMet	70S ribosome	98	1.80	0.46	44.24
2v48 W	3.80	tRNA_fMet	70S ribosome	96	1.86	0.46	44.87
3tra A	3.00	tRNA_Asp	uncomplexed	93	1.75	0.45	41.92
4tna A	2.50	tRNA_Phe	uncomplexed	100	1.70	0.52	49.00

Table 3.4: tRNA search result table (part 2).

faster than CE, and has the best characteristics regarding the TOP 10 hits in the field when considering a TM-score between 0.0 and 0.5. LaJolla's coverage and sensitivity can be adjusted using the n-gram size. The TOP 1 hits with n-gram length 10 are equally good as the results produced by TM-align. The results of LaJolla show that a certain amount of chain gets lost when using longer n-gram sizes as they cannot be indexed. As LaJolla was used with standard parameters results below a TM-score of 0.2 were neglected what also contributes to this. However, as the performance graph shows (3.8), this is a tradeoff between speed and coverage / precision. The performance is higher compared to CE and TM-align in all n-gram sizes except 10 where TM-align is faster.

3.2.4 Discussion

General aspects

The aims of this approach as defined in the introduction were the proposition of a generalized methodology that can be extended and customized by the user for different macromolecules and applications. In the results section it was shown that the performance and precision / coverage of the approach is comparable to common methods available freely today. The trade-off between performance and precision / coverage can be adjusted using the n-gram length. The described chain to string translators are independent from an initial precomputation of the secondary structure elements. With the dataset CATH_1258 derived from the 35% filtered CATH it becomes clear, that the approach works well when the sequence of the proteins is not entirely similar. Moreover, because this approach is implemented as open source in the framework LaJolla, it can be easily extended with novel translators that abstract the macromolecular structure in different ways such as suite codes or protein building blocks. It has to be pointed out that the results presented for both proteins and RNA were achieved by the backbone information alone. It is safe to assume that on both sides the accuracy of the approach could be improved by including sequence-specific information. In the case of RNA, this could be for instance the isostericity matrices of Leontis and Westhof [Stombaugh et al., 2009]. The principal performance bottleneck is the refinement step where the biomolecules have to be read from the hard disk and superimposed in 3D. Almost 80% of the time currently used for search are input / output operations. It is possible to tackle this problem from many sides. The implementation of a caching infrastructure that stores frequently used structures in memory so that subsequent hard disk reads are redirected to memory would be the first logical step. Another possibility is to store specially prepared files that only contain

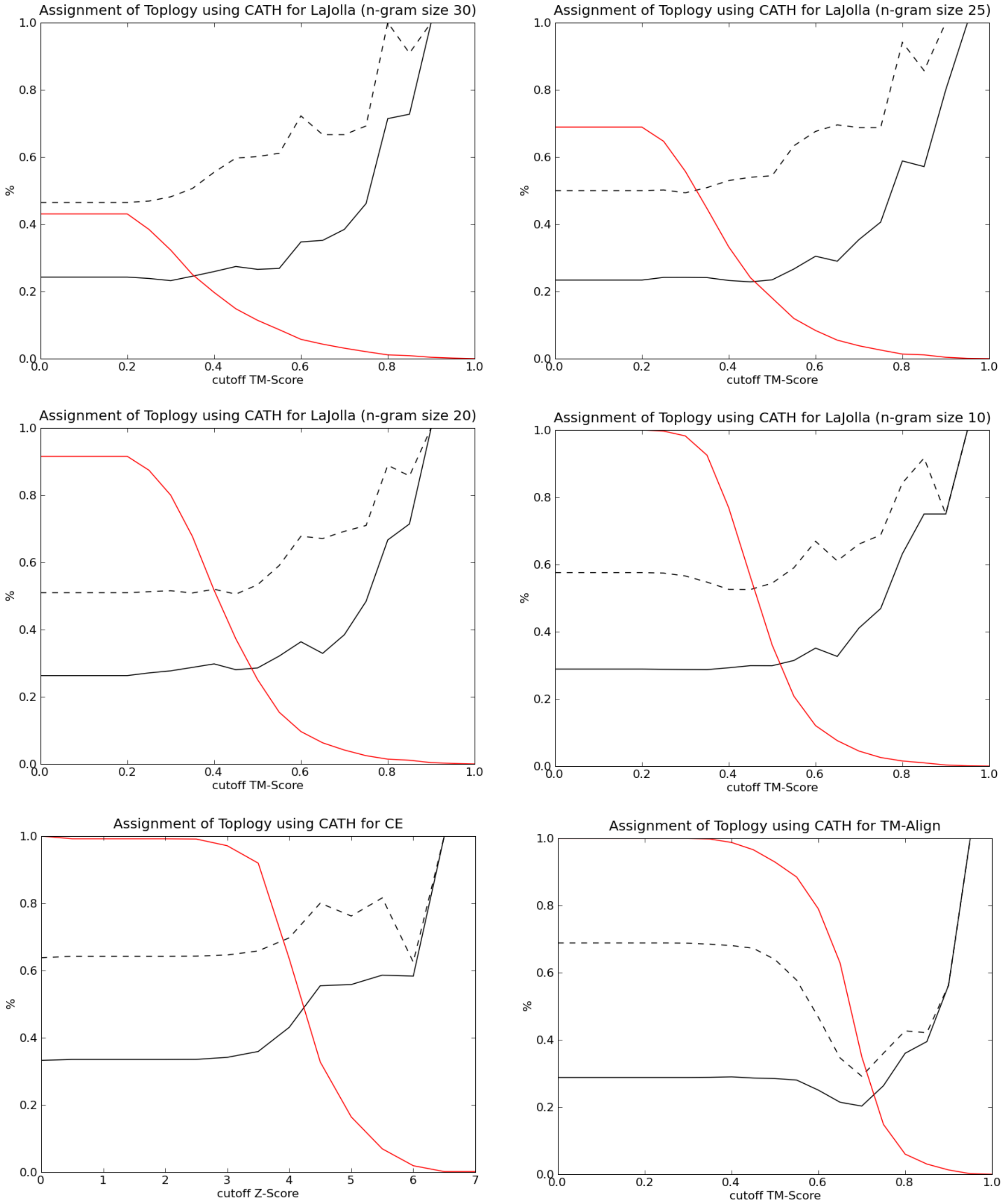


Figure 3.9: Evaluation of the coverage and precision of LaJolla (n-gram size 10, 20, 25, 30), CE and TM-align in a classification scenario. The red line indicates the percentage of the coverage of distinct topologies at a certain score cutoff. The black line represents the percentage of correct hits with the best score (TOP 1), the dashed black line represents the percentage of correct assignments with a true result being among the ten best hits (TOP 10).

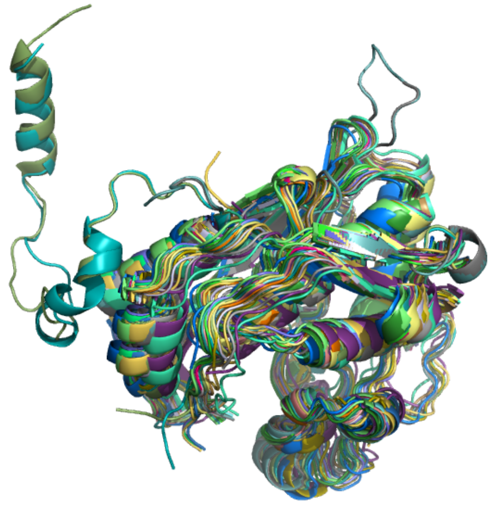


Figure 3.10: A multiple alignment of all available Thymidylate Synthases using 2tsc chain B as pivot structure.

atoms used for 3D refinement, which would reduce the file size that has to be read. Defining a threshold of how many matching n-grams between query and target at least have to be found to carry out the expensive 3D alignment has the potential to eliminate impossible alignments beforehand. Although the method was not planned to be used as tool for multiple alignment it can be used for this purpose, by the simple fact that the query structure is never translated / rotated. Subsequently, all target structures are superimposed in a multiple alignment fashion (Figure 3.10).

RNA specific aspects

The sensitivity of tRNA structural alignments is satisfactory (90.5%). In most cases, where the alignment fails, this is due to drastic structural differences, for instance in the case of lambda-tRNA, where an entire arm of the tertiary structure is displaced by an enzyme. A careful refinement of the parameters (n-gram size, TM-score threshold for alignment) could gain a few percent and superimpose a few additional examples successfully. More worthwhile to try is to run the algorithm on a vast set of RNA structures elucidating how well smaller and bigger types of RNA can be recovered. Such a study should answer how accurate the function of RNA can be recognized in general. A prerequisite for this is a careful and complete functional characterization of RNA structures that does not exist at present. Further, it could be examined whether choosing a different string representation (e.g. Richardson's suite codes) could accelerate the alignment process. But in order to not losing too much sensitivity, the n-gram search would need to account for partial similarity instead of using dissimilarity of two characters as an absolute exclusion criterion. For such and related studies, the tRNA dataset presented here provides a reasonable benchmark that could be used to compare structural search and alignment methods for RNA.

Protein specific aspects

Using the CATH as standard of truth is generally disputed. TM-align and other algorithms [Zhang and Skolnick, 2005,Pandit and Skolnick, 2008] that are originating from the field of protein structure prediction try to score a method based on the coverage of the sequence. This omits the problem that man made classification schemes such as CATH may contain wrong classifications. However, as LaJolla works completely sequence independent it is not easy to translate the meaning of the results. The CATH classification approach was used as used by other contributions [Novotny et al., 2004]. As the results of LaJolla are compared to CE and TM-align this gives a good general view of the capabilities, strengths and weaknesses of the algorithms as possibly wrong classifications are a problem for all algorithms. This methodology also allows the user to judge how to treat results with a certain score. Another general problem is that TM-align and CE do not write out protein structure positions on hard-disk by default. As LaJolla by default always writes superpositions to the hard-disk this is a clear disadvantage for LaJolla and turning off that feature would increase the performance. Still, LaJolla ranks almost always best in terms of performance even with this disadvantage. This suggests that LaJolla is especially useful when it comes to high throughput experiments, where thousands of proteins should be classified and a certain loss of coverage is regrettable.

3.2.5 Conclusions

A generalized approach for the fast search and structural alignment of arbitrary macromolecules is presented. The notion of using an index and performing one-against-all searches is a novelty in the world of RNA. This paper showed that the approach yields structural alignments that agree with biological reality using simple $\phi - \psi$ / $\eta - \theta$ translators. The described approach has an adjustable coverage and precision based on the desired speed using the n-gram size as parameter. This method will be an important aid in the high throughput functional annotation of proteins and RNA, and will make it feasible to search and test new hypotheses about protein and RNA function in a fast manner. The method has obvious applications to the field of knowledge-based docking of small compounds or even proteins. The implementation of this approach, LaJolla, is easy to extend using custom translators (eg. pure amino acid or nucleic acid sequence-based translators). The authors gladly welcome any recommendations and critiques from the community. LaJolla (including platform-independent binary packages, general development resources, mailing lists) is freely available from: http://lajolla.sf.net.

3.3 A novel approach for the detection of structural features of RNA

With the world of RNA becoming more and more important, the structural features of RNA are moving into the spotlight. The structural features of RNA long received only minor attention in comparison to the structural features of proteins. However, the rate of RNA structures being deposited into public structure archives has steadily increased to about 600 RNA structures and 1,300 protein-RNA complexes in the first quarter 2009. The evaluation of this structural space cannot be done manually, but must be carried out using computational methods. In the world of proteins, structural alignment has been state of the art since the late 1990s and a broad range of methods exist. Structural alignment of RNA is relatively new and only a few methods are available. In the following, a novel method, LaJolla2RNA, is presented. LaJolla2RNA is a novel method for aligning RNA molecules based on their $\eta - \theta$ angles. It is demonstrated that the described method provides a better overall precision and coverage in functional classification tasks compared to leading methods. LaJolla2RNA also successfully computes valid results when using known, difficult RNA alignment examples discussed in the literature. The method is capable of performing multiple alignments and is able to align large structures such as ribosomes. LaJolla2RNA (including platform independent binary package) is available as open source software from http://lajolla.sf.net. For instant access to LaJolla2RNA, a web-interface is available at http://bioinformatics.charite.de/superrnaalign.

3.3.1 Introduction

The last decade saw a growing number of research results revealing that RNA is important in key processes inside a cell and is often not translated into proteins [Eddy, 2001, Storz, 2002, Laederach, 2007, Capriotti and Marti-Renom, 2008]. It became clear that RNA is involved in post transcriptional regulation (gene silencing) via microRNAs and small interfering RNAs (siRNA) [Lim et al., 2003, Ender et al., 2008, Xiao and Rajewsky, 2009]. It was also revealed that the translational apparatus is influenced by allosteric conformational changes in riboswitches as well as frameshifts by pseudoknots and slippery sequences [Winkler et al., 2002, Penchovsky and Breaker, 2005]. RNA is furthermore involved in the chemical modification of the ribosome [Bekaert et al., 2003] and is even a player in the formation of peptides, and therefore also important for the production of proteins [Weinger et al., 2004, Nissen et al., 2000]. RNA is also important in pathological processes like cancer and retroviral infections such as AIDS [Medzhitov and Littman, 2008]. RNA is able to form complex 3D structures which are mediated primarily by hydrogen bonds formed between base pairs as well as base stacking because of its single stranded nature. The primary datasource for 3D structures of biomolecules is the PDB [Berman et al., 2007] and the number of RNA structures known and being deployed as 3D coordinates in the PDB has grown quickly in the last few years. The 3D structure of biochemical elements such as RNAs is often more conserved than its sequence, therefore the structural analysis of RNA becomes increasingly important. Another important fact is that RNA structures contain pseudo nucleotides which render it impossible to create an alignment based on the pure nucleotide sequence. In the field of proteins there are a variety of alignment and comparison techniques [Kolodny et al., 2005] able to cover a wide range of applications. In contrast, the field of RNA alignment is only now emerging, with important players being ARTS [Dror et al., 2006], DIAL [Ferrè et al., 2007], SARSA [Chang et al., 2008] and SARA [Capriotti and Marti-Renom, 2008]. In the world of proteins as well in the world of RNA there exist dictionaries with classifications of structures [Andreeva et al., 2004, Greene et al., 2007, Tamura et al., 2004]. These structural classifications can be useful when similarities between macromolecules are detected and, thus, a possibly unknown function can be inferred. In the following, a novel method is proposed: LaJolla2RNA. LaJolla2RNA is a fast and robust method for the alignment of many RNA structures. An important fact of LaJolla2RNA is that it is an open source project and designed from scratch to be modular and test-driven. This means that main parts of the framework, including e.g. the scoring function can be easily changed and monitored regarding their impact. This modular approach is, on one hand, useful when the user needs custom improvements (different scoring etc), on the other hand, LaJolla2RNA can be of interest when used in educational scenarios where students can try their own extensions to get an idea of common problems of structural alignment. For the evaluation of the method the following approach was chosen: First, the general capabilities LaJolla2RNA are examined in a functional classification scenario based on the SCOR classification (Section 3.3.3). The second part uses known difficult examples and evaluates how well LaJolla2RNA is able to solve them (Section 3.3.3). In both parts datasets and examples discussed in the literature were used to maintain a maximum of comparability to previous research.

3.3.2 Material and methods

Material

For a general overview how LaJolla2RNA performs in a functional classification scenario, two datasets of [Capriotti and Marti-Renom, 2008] are used. For a more fine-grained view all individual examples from [Ferrè et al., 2007, Chang et al., 2008, Capriotti and Marti-Renom, 2008] are applied.

trna All molecular structures containing a tRNA were retrieved from the wwPDB database [Berman et al., 2007]. The dataset was filtered manually to identify the polymer chains, to identify the functional state of the molecules and exclude structural fragments. The resulting dataset contains 101 nucleic acid chains, all of which have been resolved by X-Ray crystallography.

NR-95 The NR-95 dataset from [Capriotti and Marti-Renom, 2008] is a reduced representation of the structural universe as of November 2006. The reduction was carried out by removing RNAs with a sequence similarity of more 95%, removing structures larger than 320 nucleotides and smaller than 20 nucleotides and omitting RNA structures with only P trace atoms. The NR95 set contains 277 chains.

NR-95 SCOR The NR-95 SCOR from [Capriotti and Marti-Renom, 2008] is a subset of structures in the NR-95 set that were in the same SCOR [Tamura et al., 2004] class of functional annotations. This set contains 60 structures and 18 SCOR functional classes.

Difficult Examples The examples are made up of 7 pairwise superpositions (global, semiglobal, local) and 2 multiple alignments (tRNA and pseudoknots). Example 1 is from [Ferrè et al., 2007] example 2 from [Capriotti and Marti-Renom, 2008] and the remaining examples are taken from [Chang et al., 2008].

1. 1aszR 2csx (Note: Original 4trn replaced by 2csx)
2. Sarcin/ricin domain 28S rRNAand a 5S Ribosomal RNA (1q96A, 1un6E)
3. Pairwise global (1u8d, 1y26)
4. Pairwise global (1u8d, 1y26X:25-72)
5. Pairwise semiglobal (1hr2, 1j5a)
6. Pairwise local (1u8dA, 1y26X:39-45)
7. Pseudoknots (1l2x:A, 2a43:A)
8. tRNA (1h4sT, 1aszR:620-660, 1il2C, 2csxC, 1evvA,1j2bC)
9. Pseudoknots (1l2xA, 2ap5A, 1kpyA, 2ap0A, 1yg4A)

Method

A general introduction to the LaJolla framework is given in Section 3.2. The basis for LaJolla2RNA is the usage of transformers that translate a chain into a sequence of characters. In the case of LaJolla2RNA these characters are generated by calculating the $\eta - \theta$ pseudo-angles and translating them into characters by a discrete function. The translator *OptimizedStructureToEta-ThetaCharacterTransformer*, applied throughout this work, uses 4 discrete characters. The resulting string is subsequently chopped into n-grams of a certain size. The n-grams are in turn stored in a hash-table as keys. The corresponding values are the name of the PDB file and the position where this n-gram occurs in the PDB file. The hash-table approach has been chosen because it is a fast method to query many structures simultaneously. However, LaJolla2RNA can be used in pairwise alignments as well. If a structure (the query) is searched against one or more indexed (target) structures, the first step is to translate the query into a string. Each individual n-gram of the query string is searched in the hash-table, and a superposition based on the two matching n-grams is carried out [Kabsch, 1976]. After the superposition is finished a score based on the TM-score is calculated [Zhang and Skolnick, 2007]. The TM-score is taking into account the RMSD as well as the number of aligned residues. The scoring is carried out using the smaller of the two superimposed structures as query structure for the score by default. The best match of each pair is subsequently written out as superposition. LaJolla2RNA is based on the modular framework LaJolla. Therefore, all parts of LaJolla2RNA, e.g. scoring functions and structure to string translators, can be easily adjusted to fit the user's needs.

3.3.3 Results

The greater picture

The first objective of our research was to investigate the influence of n-gram size on the scoring function. To this end, the tRNA dataset, the complete NR-95 dataset and the SCOR subset of NR-95 were used. The structures of each dataset were aligned against each other in the same dataset. The tRNA dataset is somewhat special, as it is known beforehand that all structures have a similar shape overall. This allows to draw conclusions about how the n-gram size affects the results in recognizing similar structures. The NR-95 and the NR-95 SCOR consist of a variety of structures of all sizes.

Figure 3.11 shows that the highly similar structures of tRNA are recovered with an average TM-score of 0.45 and n-gram sizes of 5 and 10. However, using n-gram sizes 15 and 20, the average score and thus the alignment quality of the tRNA dataset declines to a TM-score of 0.37. The NR-95 and NR-95 SCOR datasets, on the other hand, behave quite similar regarding their TM-score at different n-gram sizes. The average TM-score improves when the n-gram size is increased. As a certain amount of structures are smaller than the n-gram size it is obvious that some structures cannot be found with large n-gram sizes. To investigate into this, the number of recovered tRNA structures using different n-gram sizes is shown. The tRNA dataset is useful for this, as it is known that each alignment is valid beforehand, a fact not known for the NR-95 and the NR-95 SCOR datasets. Figure 3.12 shows that with n-gram sizes 5, 10 and 15 almost all tRNA structures are found, whereas, with n-gram size 20, only 75 % of tRNA structures are found. At the same time (as shown in Figure 3.11), the TM-score of the tRNA structures using n-gram size 20 was significantly lower than using n-gram size 5, 10 or 15.

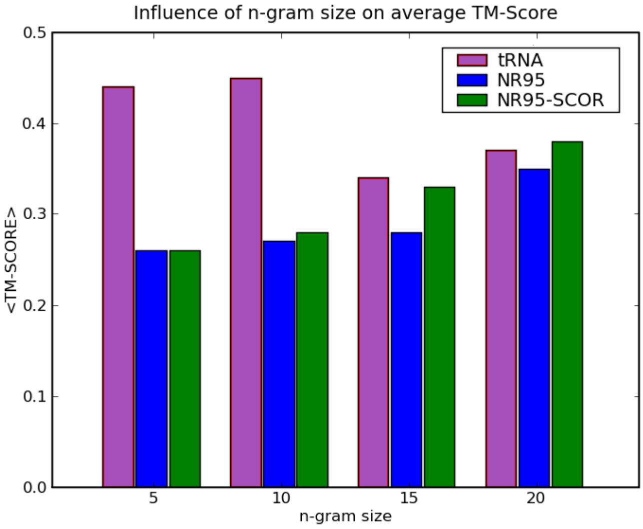


Figure 3.11: Average TM-score of all against all searches with datasets tRNA, NR-95 and NR-95 SCOR using different n-gram sizes.

To investigate how well LaJolla2RNA is able to cope with a classification task, the NR-95 SCOR dataset was used. This dataset contains only structures that are annotated using the SCOR functional classification. Therefore meaningful results can be detected by comparing if the results retrieved by LaJolla2RNA have the same SCOR classification as the query structure. All structures were aligned against all structures and it was checked whether the best hit (TOP 1) was a true positive or if a true positive was among the best five hits (TOP 5). Figure 3.13 shows the results using n-gram size 10 against

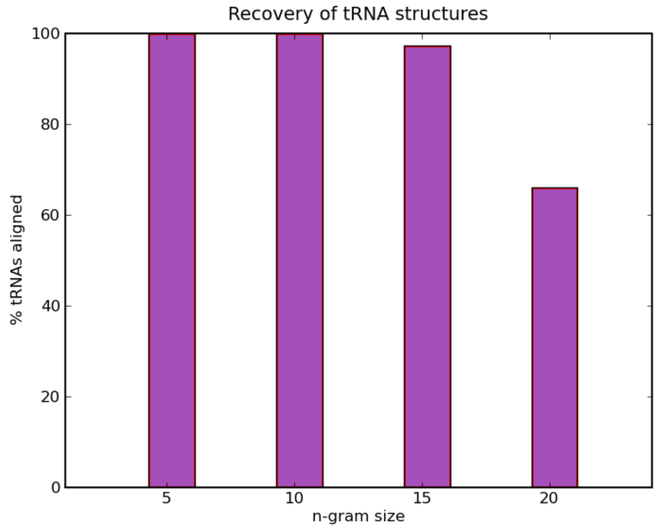


Figure 3.12: Recovery of structures in the tRNA dataset using different n-gram sizes.

the coverage and different scoring cutoffs. As the same dataset, proposed by [Capriotti and Marti-Renom, 2008] is used, the results can be compared directly to the method SARA. LaJolla2RNA shows a better characteristic considering the coverage of SCOR functional classes over almost the entire spectrum. At the intersection between coverage and TOP 1 results, LaJolla2RNA is almost 10% better than SARA. However, it must be noted that LaJolla2RNA does miss some structures (maximum coverage is 95%) with n-gram size 10 that are recovered by SARA.

Assessment with difficult examples

The overall picture as presented in Figure 3.13 shows that the approach is able to cope with leading methods in functional classification tasks. However, this does not necessarily mean that the approach solves difficult RNA alignment problems. Each of the former contributions to the field of RNA alignment therefore shows at least one example where one algorithm is better than another. All examples were collected from [Ferrè et al., 2007,Chang et al., 2008,Capriotti and Marti-Renom, 2008] in dataset difficult_examples (see Section 3.3.2) and checked whether LaJolla2RNA was able to compute the correct solution. The conclusion is that LaJolla2RNA successfully solves all alignment examples where two structures are aligned against each other. LaJolla2RNA solves them with standard parameters and n-gram size 5. However, the two

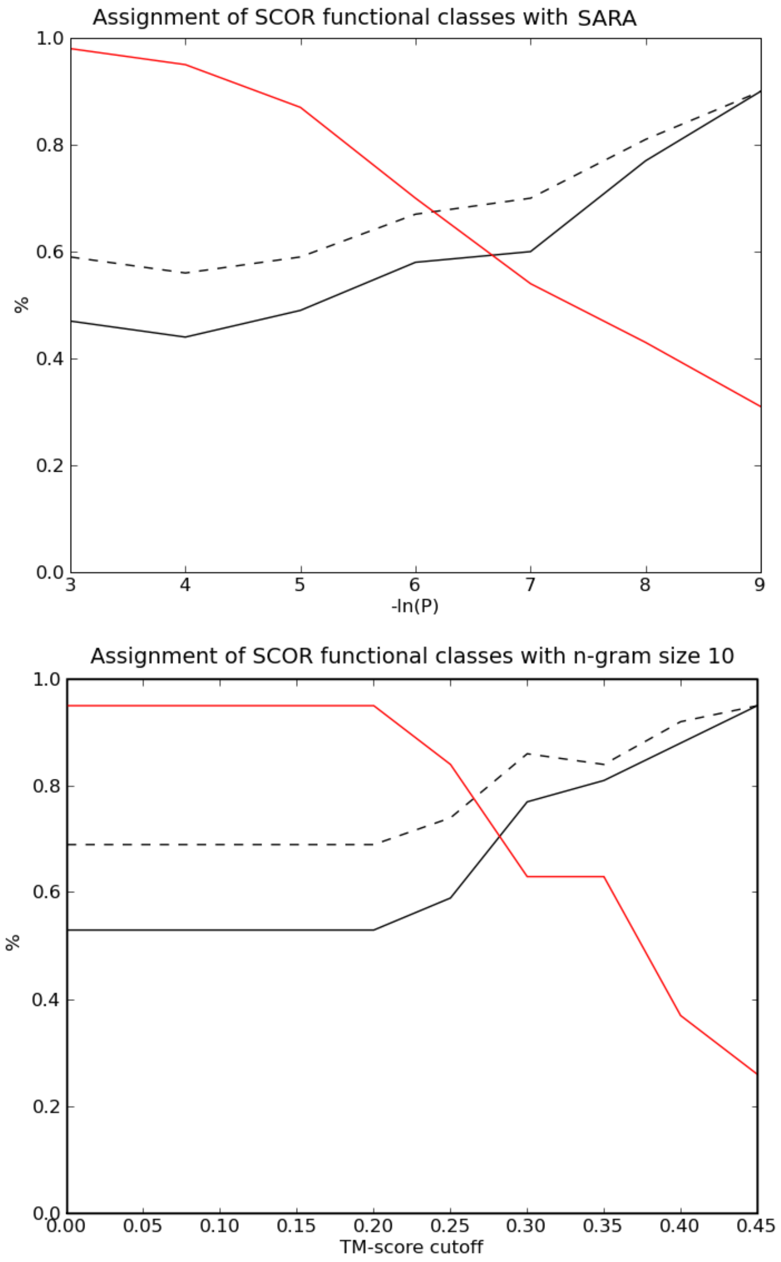


Figure 3.13: Evaluation of the coverage and precision of SARA [Capriotti and Marti-Renom, 2008] (top) and LaJolla2RNA, n-gram size 10 (bottom) in a classification scenario. The red line indicates the percentage of the coverage of distinct SCOR functional classes at a certain score cutoff. The black line represents the percentage of correct hits with the best score (TOP 1), the dashed black line represents the percentage of correct assignments with a true result being among the five best hits (TOP 5).

multiple alignment examples can only be solved when a good pivot structure is chosen. When a pivot structure is chosen that is dissimilar, compared to the rest of the ensemble, the multiple alignment does not work well. Using the most similar structure in the ensemble LaJolla2RNA also solves all multiple alignment tasks successfully.

In addition, LaJolla2RNA is also able to superimpose large structures and is able to generate multiple alignments of a vast amount of structures. To this end, two ribosomes are aligned (Figure 3.14) as well as all structures in the dataset tRNA (Figure 3.15).

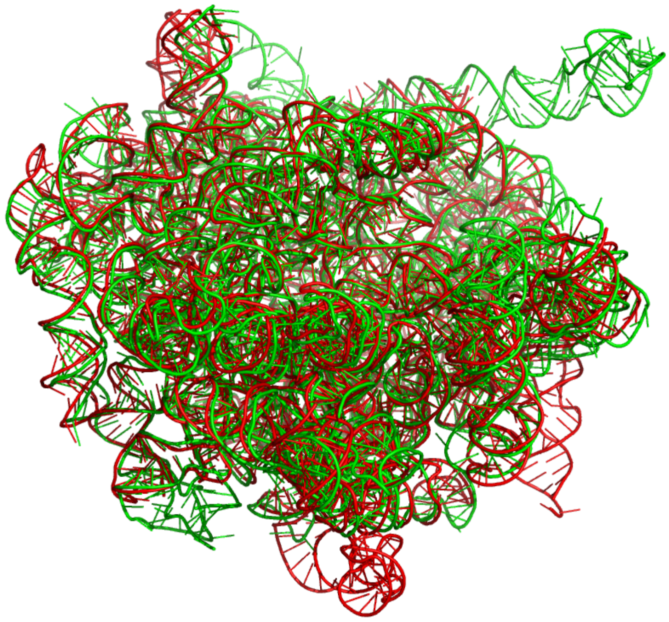


Figure 3.14: Example of LaJolla2RNA aligning two large ribosomes (red: 1ffk, green: 2v47).

3.3.4 Discussion

In a functional classification scenario, LaJolla2RNA outperforms SARA in both TOP1 and TOP5 hits given a certain coverage. The difficult pairwise alignment examples presented by three other methods are also solved successfully by LaJolla2RNA. This is encouraging, as each of the examples has been proposed by the original method to show its superior qualities in comparison to another method. LaJolla2RNA solves all alignment tasks using standard parameters. However, regarding the multiple alignment, there is one drawback.

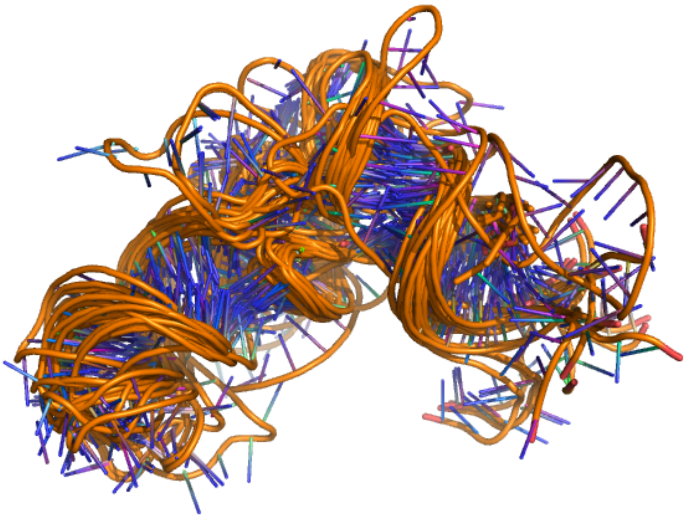


Figure 3.15: A multiple alignment of all structures in the dataset trna using 1b23 as pivot structure.

LaJolla2RNA in the current configuration only generates multiple alignments based on a pivot structure. If the most dissimilar structure in an ensemble is chosen, there is a good chance that the multiple alignment will not work for all aligned structures. This can be solved in the future by an exhaustive procedure where subsequently all structures are chosen as pivot structures and only the best result is kept. This is, however, not implemented in the current version. It is also a matter of opinion whether it makes sense to use the TM-score for RNA structures as the TM-score was originally developed for protein alignments. Our experiments suggest that TM-score is a good choice, but it becomes clear that the full range of the score (0,1] is not utilized. For instance a score above 0.45 only rarely occurs, even when the structures are very similar, as demonstrated in the trna dataset. This can potentially be corrected by adjusting the parameters of the equation. Depending on the intended use, the n-gram size is very important. When the n-gram size is too big, smaller structures cannot be found, simply because they are not stored in the hash-table of the algorithm. Simply using small n-grams is not a solution, because smaller n-gram sizes make the procedure slower and also introduce false positives to the results.

3.3.5 Conclusions

LaJolla2RNA has been shown to deliver good results in functional classification tasks. It also delivers good results when assessing the method using difficult examples in pairwise as well as multiple alignments. As the framework is available as open source, each aspect discussed can be adjusted according to the user's needs. LaJolla2RNA can also be of special interest in a teaching environment. LaJolla2RNA (including platform independent binary package) is available as open source from http://lajolla.sf.net. For instant access to LaJolla2RNA a web-interface is available at http://bioinformatics.charite.de/superrnaalign.

Chapter 4

Simulation of cellular reactions

4.1 Quantitative simulation of apoptosis using Petri nets

This section is driven by the fact that the genetic and regulatory mechanisms of apoptosis (the programmed cell death) are key players in several diseases. In the following, a concept for quantitative simulation of apoptosis using Petri nets is evaluated. The term “quantitative” is defined by the usage of microarray or proteomics data for the signal transduction of the net. The simulations presented in Section 4.1.4 were carried out by Christian Scholz as part of his master’s thesis [Scholz, 2008].

4.1.1 Introduction

Life is about interaction and regulation which can be modeled by biological networks. This modeling can be carried out on different levels, e.g. metabolic pathways or signal transduction networks. Metabolic networks describe the flow of metabolites caused by enzymatic reactions using their stochiometric properties, e.g. the glycolysis metabolism or pentose pathway. Signal transduction networks consist of cascades of proteins and protein complexes. They describe how cells process internal as well as external signals such as changes in phosphorylation, ion channel activity and gene expression.

In theory, a cell consists of various pathways that are often defined by a certain theme such as apoptosis. Combination of pathways leads to a more complete view and model of a cell. An aim is to simulate perturbations by the knockout of interaction partners. These perturbations can lead to a better understanding of how drugs affect cells and to the directed development of novel drugs. Often these perturbations can only be accomplished by time consuming *in vitro* experiments. Therefore, *in silico* simulation is of great interest for the elucidation of network properties as well as the directed development of treatments. A problem in pathway creation is that relevant information is often distributed over various databases and repositories. Often pathways are discussed controversially and sometimes contradict each other.

4.1.2 Petri nets in systems biology

Petri nets are a modeling method that allow for an extensive model evaluation based on formal semantics. Hence, they are descriptive and easy to read at the same time [Reisig, 1985]. Carl Adam Petri developed Petri nets as pragmatic concept to write down basic chemical reactions during his high school years (Figure 4.2). In 1961 he submitted his dissertation in computer science where a formal concept for the simulation of distributed and concurrent systems is presented. Petri nets can be used to model qualitative as well as quantitative aspects of biological systems [Lee et al., 2004, Heiner et al., 2004, Peleg et al., 2005, Sackmann et al., 2006, Grunwald et al., 2008]. The core concept can be extended as shown by e.g. colored Petri nets, stochastic Petri nets and hybrid Petri nets [Lee et al., 2006, Chaouiya, 2007].

Apoptosis

Apoptosis, the programmed cell death, is a feature of multi cellular organisms [Elmore, 2007]. It involves a range of biochemical events such as changes in morphology, changes to the cell membrane, blebbing, shrinkage, nuclear fragmentation, chromatin condensation, and chromosomal as well as DNA fragmentation. The programmed cell death is an important concept in many regards, such as fetal development and the development of the nervous system and the way organism deal with pathologic cells. Apoptosis is important for the understanding of various diseases, e.g. AIDS, cancer, and Alzheimer as well as Parkinson's disease [Chu and Chen, 2008, Storey, 2008, Borden et al., 2008].

The programmed cell death is regulated by pro-apoptotic and anti-apoptotic factors. An important player in the apoptosis pathways are caspases, a class of cysteine proteases [Van De Water et al., 2004]. Upon activation by proteolytic cleavage, caspases degrade cellular targets. Two pathways can lead to the activation of caspases: The extrinsic and the intrinsic pathway. The extrinsic pathway is activated by ligands that bind to death receptors e.g. TNFR-1, Fas and the glucocorticoid receptor. The intrinsic pathway is activated in response to e.g. growth factor deficiency and DNA damage. Apoptotic stimuli can change the permeability of the outer mitrochondrial membrane and cause

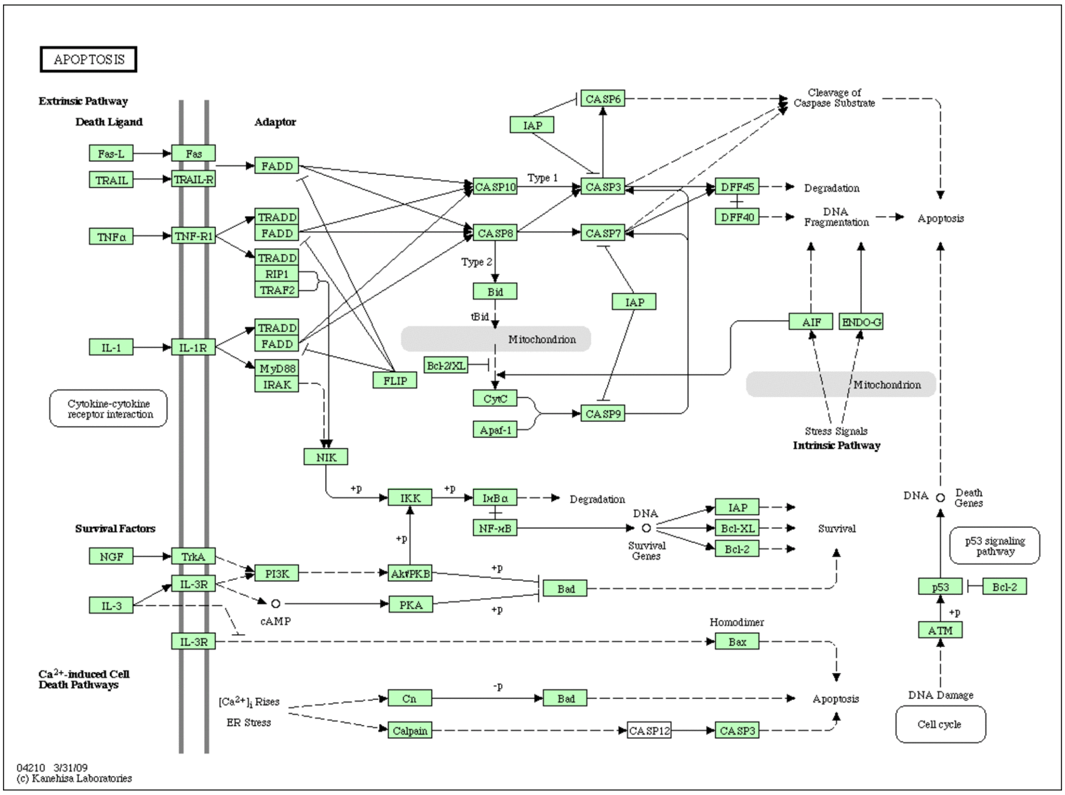


Figure 4.1: The apoptosis pathway for homo sapiens as described by KEGG (accession id hsa04210). KEGG gives an informal overview about the interaction partners (proteins, rectangles), the interactions (arrows) and the outcomes (described in natural language such as apoptosis, and survival).

cytochrome c release and other proteins to enter the cytosol. Cytochrome c then binds to caspase activating proteins such as Apaf-1 and pro-caspase-9. The release of cytochrome c is mediated by proteins of the Bcl-2 family. The major anti-apoptotic proteins are Bcl-2 and Bcl-XL, the major pro-apoptotic proteins are Bax, Bad, Bim, and Bid [Huang and Strasser, 2000, Flemming, 2008].

Goal

The first goal is to generate a suitable Petri net based signal transduction model that represents the apoptosis pathway in a detailed manner. The second goal is to evaluate various possibilities of mapping microarray expression levels on that network and propose a method suitable to predict the outcome of *in vitro* experiments.

4.1.3 Material and methods

The NCI cell lines

The data used to quantify the network is originating from the National Cancer Institute (NCI). The NCI is testing constantly how cells react to the application of various compounds. To this end, the NCI maintains a set of cancer cell lines from different tissues. The NCI also provides RNA expression data of the cell lines for about 60 cell lines. The data used in this work is based on a Affy U133 microarray chips and freely available [Ross et al., 2000]. The cell lines used in this study are listed in Table 4.1.

Petri nets

Petri nets contain two different types of nodes: places and transitions. Commonly, places are drawn as circles, and transitions are depicted as rectangles. Places and transitions are connected via directed arcs. Arcs always connect places with transitions and never transitions with transitions or places with places. Places can contain a non-negative value of tokens. The tokens of all places are called the “marking” of a Petri net. The initial state of a Petri net is called “initial marking”. A transition consumes tokens from its input places and is producing tokens on its output places (Figure 4.2). The consumption and production of tokens by a transition is called the firing of a transition. A transition is able to fire when the places of each incoming arc are filled with at least the number of tokens labeled on the corresponding arc. If there is no label on the arcs a value of 1 is assumed. Accordingly, the transition produces as many tokens on the output places as indicated by the labels on

Tissue	Name cell line
Lung	NCI-H23, NCI-H226, NCI-H322M, NCI-H460, NCI-H522, A549-ATCC, HOP-62, HOP-92, EKVX
Ovarian	OVCAR-3, OVCAR-4, OVCAR-5, OVCAR-8, IGROV1, SK-OV-3
Central nervous system	SNB-19, SNB-75, U251, SF-268, SF-295, SF-539
Hematop	CCRF-CEM, K-562, MOLT-4, HL-60(TB), SR, RPMI-8226
Colon	HT29, HCC-2998, HCT-116, SW-620, COLO205, HCT-15, KM-15
Renal	UO-31, SN12C, A498, CAKI-1, RXF-393, ACHN, 786-0, TK-10
Melanoma	LOX_IMVI, MALME-3M, SK-MEL-2, SK-MEL-5, SK-MEL-28, UACC-62, UACC-257, M14

Table 4.1: Tissue and name of cell lines used in this study.

the individual outgoing arcs, 1 in case of no label. A Petri net can use different rules determining how the tokens flow in a net. The basic firing rule of Petri nets is non-deterministic. This means that each enabled transition is selected by chance to fire. Repeatedly firing a transition is the process that produces a simulation. An extension of the basic firing rule is the stochastic non-deterministic firing rule where the selection of the transition to fire is not distributed evenly. Each transition gets a probability where transitions with higher values are more likely to fire. A comprehensive introduction to Petri nets can be found in [Reisig, 1985].

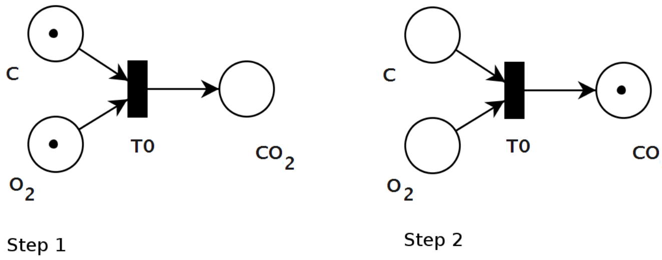


Figure 4.2: A Petri net example from the field of chemistry, where CO_2 is produced from C and O_2. In Step 1 the markings in place C and O_2 indicate that transition $T0$ is able two fire. In Step 2 transition $T0$ consumed the tokens of places C and O_2 and produced a token in place CO_2.

Mapping methods

Four methods, with the potential to map quantitative data of RNA microarrays on a signal transduction Petri net, are evaluated.

Method 1 - weighting incoming edge weights Mapping method 1 uses rounded integer values of the expression levels. The expression level of each protein was mapped on the incoming edge as weight. In case of more than one edge pointing to a place of a certain gene, the value was divided by the amount of incoming edges. The resulting value was used as incoming weight for each of them.

Method 2 - weighting incoming and outgoing edge weights Mapping method 2 uses rounded integer values from the expression levels as values for the incoming and outgoing edges. If there is more than one incoming or outgoing edge the value is evenly distributed on the edges. For instance in case of a gene having an expression level of 20 and one incoming E1 and two outgoing edges E2, E3. E1 would be 20 and E2 and E3 would be 10.

Method 3 - using an initial marking The third mapping method uses an initial marking consumed during the simulation. For the initial marking, rounded integer values of the expression levels were used. The weight of all edges is set to 1.

Method 4 - stochastic transitions and incoming edge weights Mapping method 4 uses the expression level as firing probability of a transition. The relative RNA amount of a protein with respect to all other proteins in a net was calculated. The resulting value was used as firing probability. Additionally, the maximum amount of tokens in all places except apoptosis and survival was limited to 5.

The signaling network The network was built to study the interaction of proteins that induce and inhibit apoptosis. The basic template for the network was taken from the KEGG [Kanehisa et al., 2007], depicted in (Figure 4.1). The final Petri net integrates the extrinsic and intrinsic apoptosis pathways (Figure 4.1). The extrinsic pathway was integrated using the death receptor, the intrinsic pathway is integrated via the mitrochondric signals and the proteins Bid and Bad. Bcl-2 and Bcl-XL were integrated as well as their antagonists Bax and Bak. The final Petri net contains a place for apoptosis and a place for survival. The two special places act as indicators which describe the signal flow in the network. A higher number of tokens in the survival place indicates a typical behavior of a cancer cell. Apoptosis is primarily induced by a signal from Caspase 3. Survival is induced when Bcl-2 or Bcl-XL are interacting with Bax and Bak.

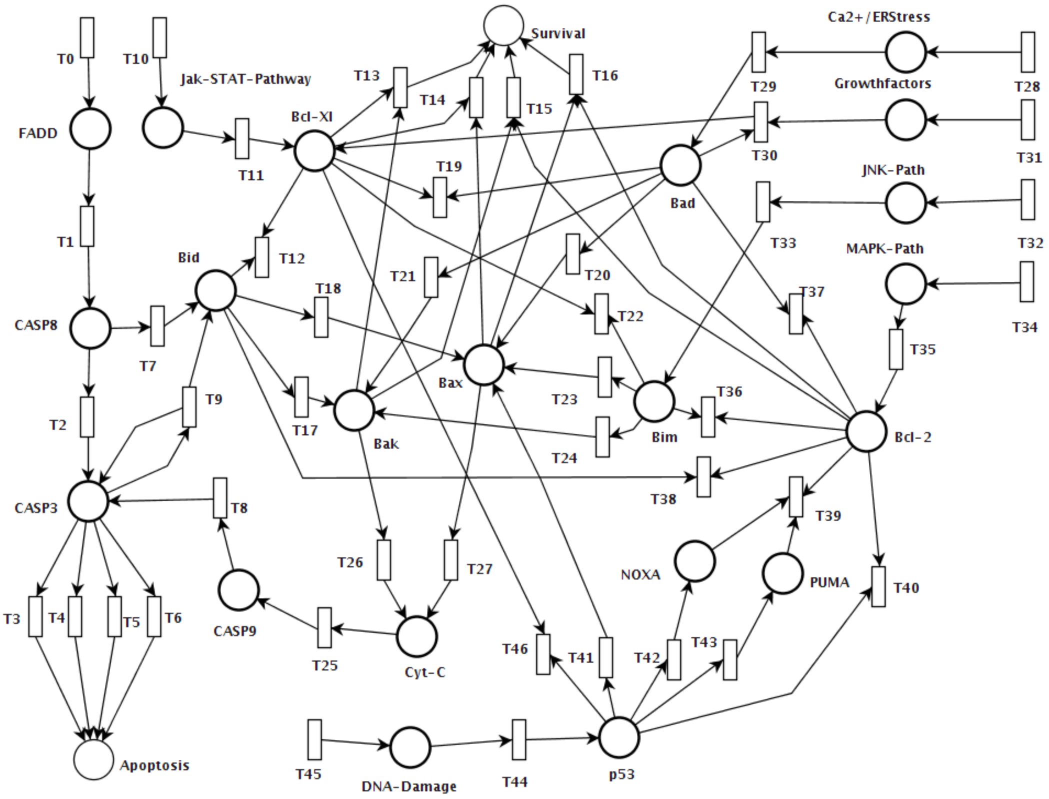


Figure 4.3: The Petri net for apoptosis used in the simulation.

4.1.4 Results

The simulation results were generated using PIPE 2.5 (http://pipe2.sf.net).

Simulation

Mapping method 1 After 10,000 simulation steps the marking of the network reflected the initial expression levels of the corresponding regions. However, the number of tokens in the apoptosis and survival places was about the same (Table 4.2). This contradicts the distribution of pro- and anti-apoptotic genes that are active in the cell lines. However, this outcome can be explained by the fact that each of the incoming transitions to apoptosis and survival is permanently active. To an extent, the final marking in the apoptosis and survival places only reflects the evenly distributed firing of the incoming transitions of those two places.

Cell line	CCRF-CEM	CAKI-1	TK-10
Apoptosis	0.50	0.46	0.51
Survival	0.50	0.54	0.49

Table 4.2: Results mapping method 1.

Mapping method 2 To reduce the amount of generated markings in the places the outgoing arcs were also weighted relative to their expression levels. The marking of the network after 10,000 simulated steps no longer corresponds to the expression levels of the relevant genes. However, like in Model 1 the marking of the two places apoptosis and survival is also evenly distributed (Table 4.3). This is again due to the fact that this distribution only reflects the firing of the incoming places.

Cell line	CCRF-CEM	CAKI-1	TK-10
Apoptosis	0.48	0.48	0.49
Survival	0.52	0.52	0.51

Table 4.3: Results mapping method 2.

Mapping method 3 Mapping method 3 uses an initial marking as starting point for the simulation. After 10,000 simulation steps 60% of the tokens were in the apoptosis pathway, 40% of the tokens were in the survival place 4.4. This was the case for all 3 cell lines (CCRF-CEM,CAKI-1 and TK-10). This concept also was not able to reflect the amount of apoptosis induced by a cell line.

Cell line	CCRF-CEM	CAKI-1	TK-10
Apoptosis	0.65	0.60	0.59
Survival	0.36	0.40	0.41

Table 4.4: Results mapping method 3.

Mapping method 4 Mapping method 4 uses the results of the previously presented mapping methods 1-3. On the one hand mapping method 4 uses the expression levels as probability for a certain incoming transition of a protein to fire. The higher the expression level the more signals are coming from a certain protein. On the other hand the previous mapping methods showed that an accumulation of tokens happens in many places. Therefore it was decided to use a limit of 5 tokens of each place in the network. This can be observed in biological cells as well, where a certain saturation of proteins takes place. After 5,000 fired transitions cell lines CAKI-1 and TK-1 accumulate more tokens in the survival places than in the apoptosis places. CCRF-CEM showed an almost even distribution of apoptotic and survival signals. This suggests that the paradigm chosen for Model 4 corresponds to the biological reality where cancer cell lines tend to suppress apoptosis (Table 4.5).

Cell line	CCRF-CEM	CAKI-1	TK-10
Apoptosis	0.52	0.18	0.19
Survival	0.48	0.82	0.81

Table 4.5: Results mapping method 4.

Mapping method 4 was subsequently applied to cell lines of the NCI set. When running the simulation the cells are supposed to have a higher amount of tokens in the survival place than in the apoptosis place. The results show that this is only true for about half of the tested cell lines (Figure 4.4). However, there is a clear correlation between the amount of apoptosis predicted and the RNA expression level of pro-apoptotic proteins in the cell lines (Figure 4.5). This is also true for the anti-apoptotic expression levels (Figure 4.6).

4.1.5 Discussion and future directions

The concept of using the expression levels as weights on the edges of the places as applied by mapping method 1 and 2 did not work well. This was due to the fact that tokens are accumulating in all places. Therefore the transitions leading to the places for apoptosis and survival were always enabled. This resulted in an even token distribution in apoptosis and survival. Mapping method 3 used the expression level as initial marking for the net. The assumption was that the network runs out of tokens and the tokens will in turn accumulate in the apoptosis or survival place. However, the places apoptosis and survival once again showed an even distribution of tokens. This was due to the fact that the input transitions were always enabled. Another reason was that the transitions coming into the network were still able to flood the net. The most

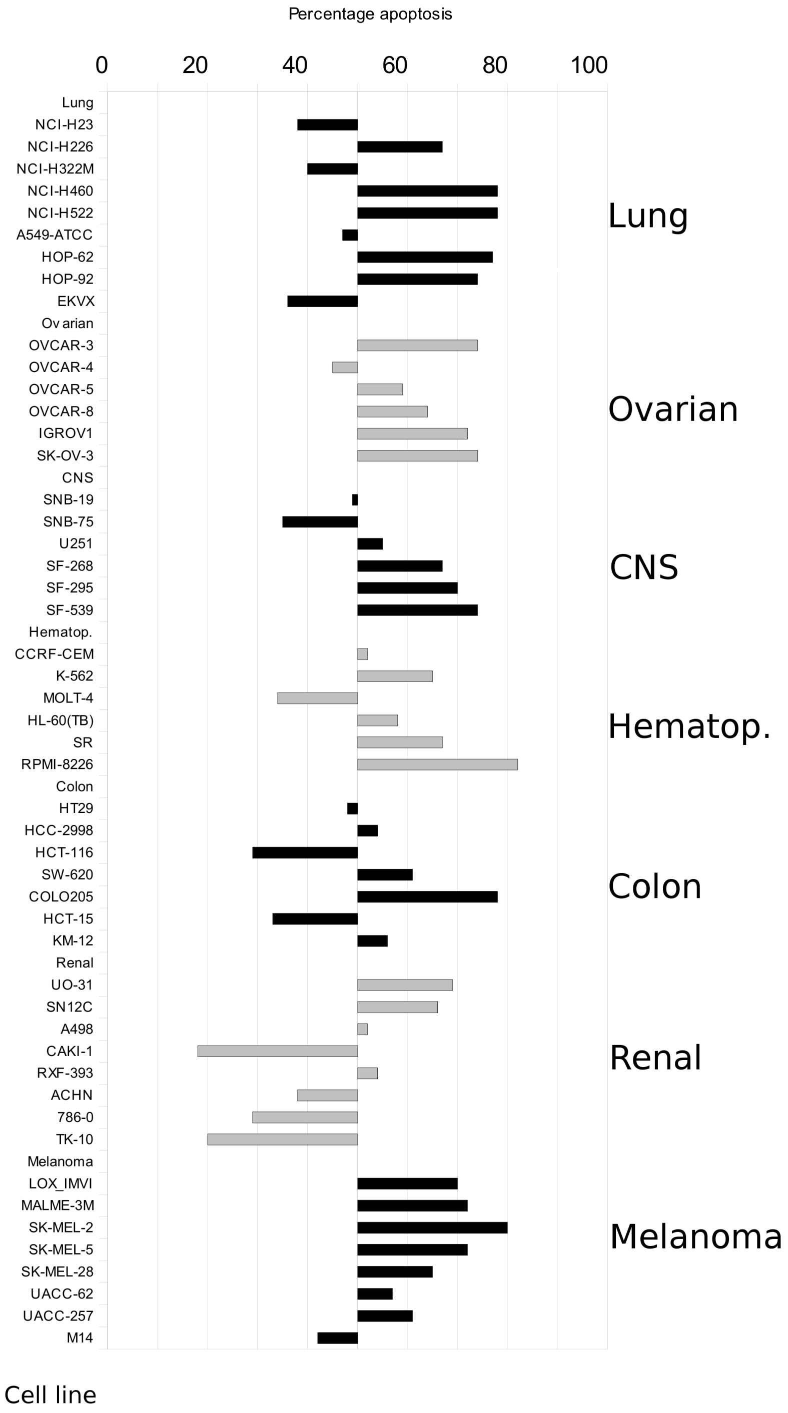


Figure 4.4: Percentage of tokens in apoptosis place using mapping method 4 on NCI cell lines.

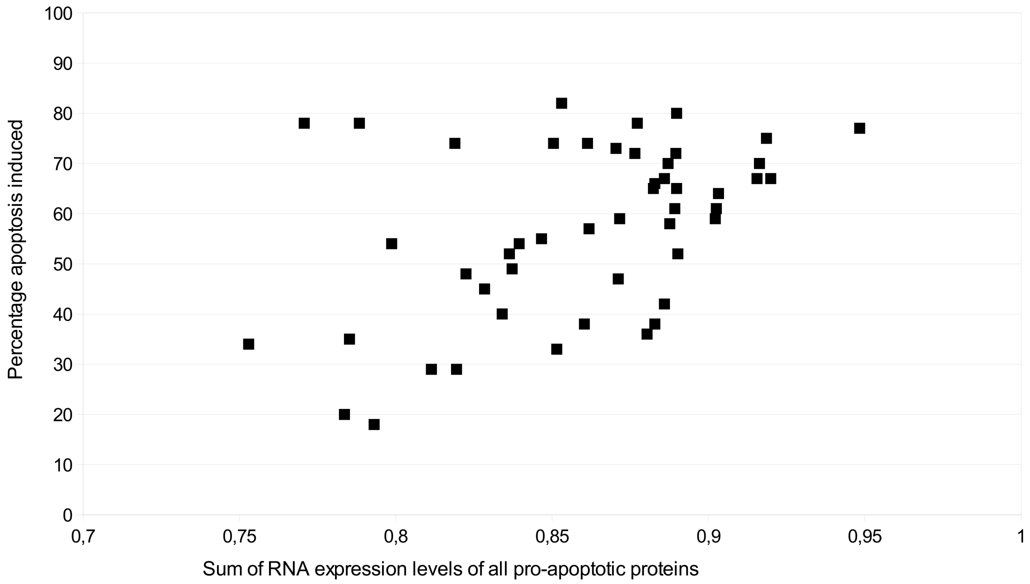


Figure 4.5: Sum of RNA expression levels of pro-apoptotic proteins of the NCI cell lines in relation to the predicted apoptosis using method 4.

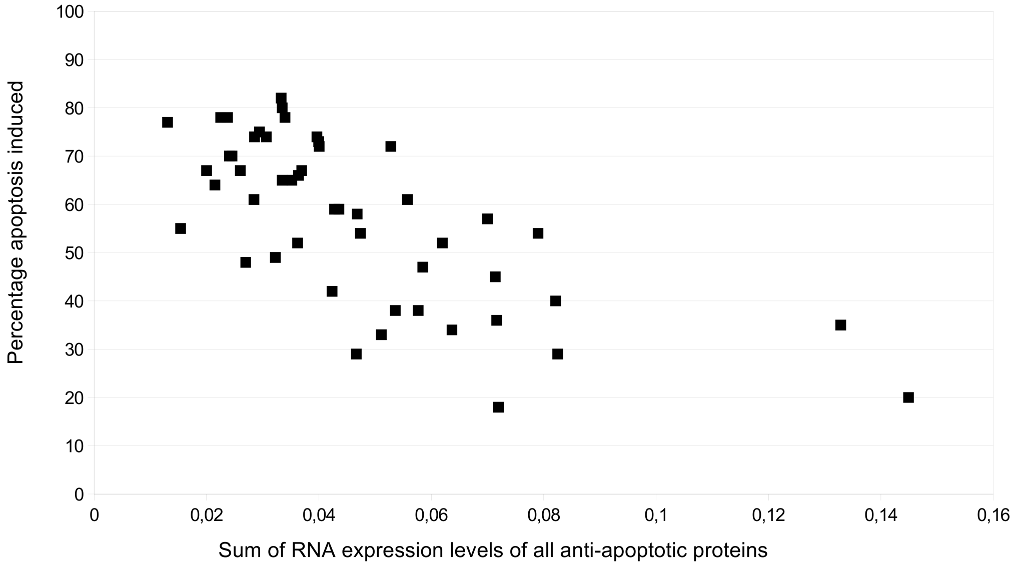


Figure 4.6: Sum of RNA expression levels of anti-apoptotic proteins of the NCI cell lines in relation to the predicted apoptosis using method 4.

promising concept was the usage of expression levels as firing probabilities for a stochastic non deterministic Petri net represented in mapping method 4. The maximum amount of tokens per place was limited to 5 what effectively stopped the flooding of the net. The results showed that the exemplary cell lines produce more survival than apoptosis signals in case of cell lines CAKI-1 and TK-10, or an almost even amount in case of cell line CCRF-CEM. When applied to NCI cell lines with microarray data available the prediction of the network is not correct in all cases. However, there is a correlation between the amount of apoptosis predicted and the RNA amount of pro-apoptotic proteins present; the same is true for anti-apoptotic proteins correlating with a higher level of survival. In cases with no correlation the cell property "cancer" may not be connected with the suppression of apoptosis. The next step is to weight the incoming transitions of the apoptosis pathway, or to extend the network to incoming signaling networks. Currently, the incoming signals of the signaling networks tend to flood the network – even in mapping method 4. To check whether the model corresponds to the biological reality knock out experiments have to be carried out *in vitro* and the results checked *in silico*. This validation should ideally be carried out using protein quantification techniques. This omits a likely bias that is introduced as not all RNA is translated into protein.

4.1.6 Conclusions

Section 4.1 presents a novel concept to map expression data on signal transduction pathways. The amount of signal transduction in a cell is dependent on the amount of protein. To this end, expression levels of genes are mapped on a Petri net using different mapping methods. For the current evaluation the apoptosis signaling pathway was chosen. The most promising method 4 uses expression levels as firing probabilities for transitions and a stochastic non-deterministic Petri Net. Model 4 showed good compliance with biological reality and is the basis of further extension of the network to more pathways. The result of a combined effort is the simulation of a complete cell.

4.2 Cell Sim – a Petri net based cell simulation software

During the work presented in chapter 4.1 it became apparent that many software solutions for simulating and analyzing Petri nets are not suitable for the life science domain. To carry out the experiments and extend the approach it was important having a simulation environment that is usable by biologists and extendable to new requirements arising from simulation results such as new firing rules. The implementation of the software solution, called Cell Sim, was carried out by Konstantin Pentchev in his bachelor's Thesis [Pentchev, 2008].

4.2.1 Introduction

There exists a variety of open source generalized Petri net simulators such as PIPE (http://pipe.sf.net), PNK (http://www2.informatik.hu-berlin.de/top/pnk/), snoopy [Heiner et al., 2008] and others. However, all generalized tools only offer the standard Petri net view that is too complex for biologists and blocks them from using the software and carrying out *in silico* experiments. On the other hand the possibilities of these simulators where not satisfying. Snoopy for instance is not able to run stochastic Petri nets, PIPE on the other hand saves files in a non standard compliant format. Concerning Petri net tools especially drafted for life sciences, Cell Designer is a prominent example. However, Cell Designer has not the range of firing rules needed for the experiments besides the fact that there was no update the last 2 years. It is also a commercial tool that cannot be modified to our needs. Therefore it was decided to develop a novel extendable Petri net simulation solution that can be used by biologists to experiment with Petri nets.

This software solution was designed with the following requirements:

- Platform independent / Startable from the web without prior installation
- Reuse as many components as possible / Be standard compliant
- Easy to understand and to use for non computer scientists

4.2.2 Material and methods

Platform independent / startable from the web

To be platform independent it was decided to use Java together with the SWING GUI library. This allows the software to run on almost all modern computers. To run the software as web application a technology called "Java

Web Start" was applied that is able to run the application without previous installation simply by clicking a button on a webpage. Often, users find it too complicated to install an application before execution. Therefore the web start possibility is important for first time users.

Reuse as many components as possible / standard compliance

Both objectives were met using the Petri net Kernel (PNK, http://www2.informatik.hu-berlin.de/top/pnk), an open source software framework developed in the lab of Wolfgang Reisig. The framework software is able to visualize a Petri net, but also has the possibility to run a simulation as well as extended possibilities to read and write Petri net files. The standard file format for Petri nets is PNML, an XML dialect supported by the PNK out of the box. The PNML compliance allows the user to use the Cell Sim software in concert with other Petri net software facilitating e.g. further analysis of the net.

Easy to understand and use for non computer scientists

Researchers working with cells in the lab are often familiar with notations used in the KEGG [Aoki and Kanehisa, 2005], but they are overwhelmed by the complexity of Petri nets. To this end, three views of a Petri net were implemented that represent the Petri net on different levels of complexity. These views are generated automatically or semi automatically. The first View (Figure 4.7) is a plain Petri net view that is easy to understand for computer scientists but often too complex for wet lab scientists. The second view is of intermediate complexity showing properties of Petri nets, but in a KEGG-like fashion (Figure 4.8). The third view is a complete abstraction from the Petri net similar to a KEGG map (Figure 4.9).

4.2.3 Results

The software was successfully tested under Linux/QT, Mac OS X 10.5 and Windows Vista.

4.2.4 Discussion and conclusions

This section presented a software tool for the generation and simulation of signal transduction networks as Petri nets. The tool is tailored towards the needs of life sciences by providing additional views that are modeled after the popular KEGG nomenclature. The software tool is built with standard compliant exchange formats such as PNML and can be extended to the user's needs. In the future it is planned to release the software as open source. The software will also be a fundamental building block for a virtual cell that will be realized with the concepts presented in Section 4.1.

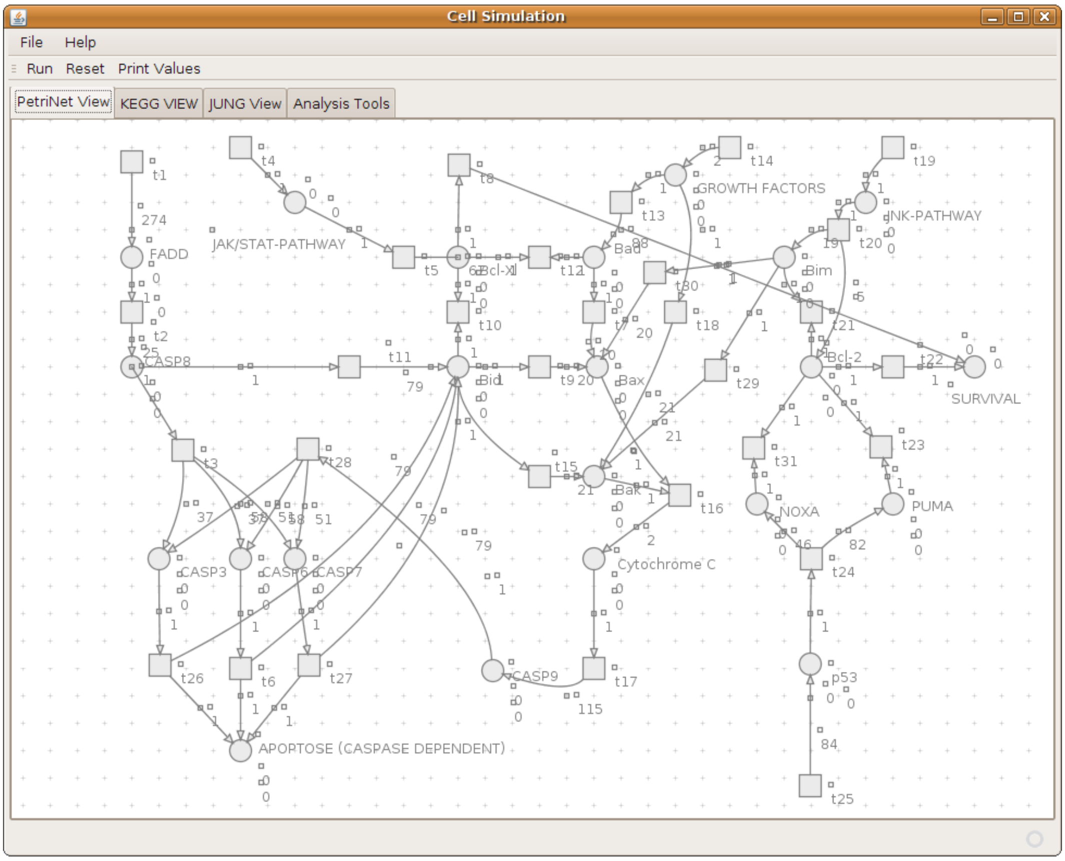


Figure 4.7: The Graphical User Interface of Cell Sim is capable of showing different levels of regular complexity of users. The screenshot shows the low level Petri net view best suited for experienced users.

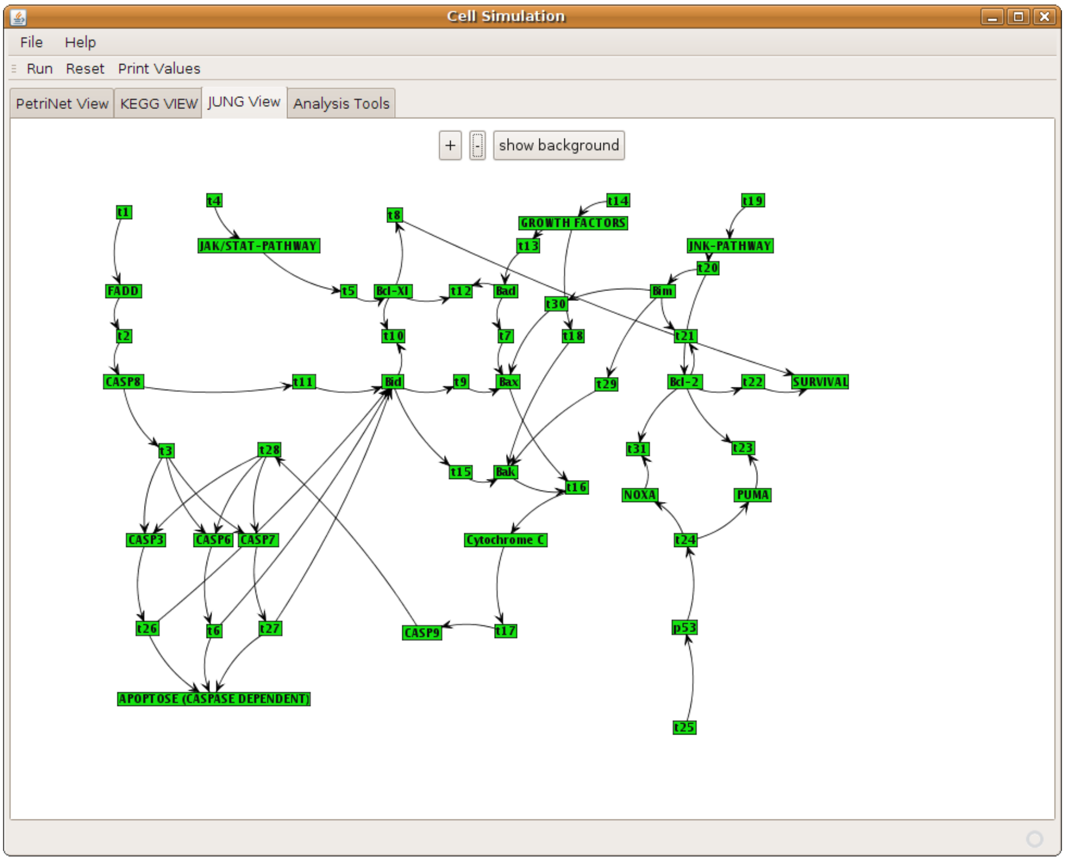


Figure 4.8: The Graphical User Interface of Cell Sim displaying the JUNG view of intermediate complexity. The view consists partly of Petri net elements and partly of KEGG style elements understandable to non Petri net experts.

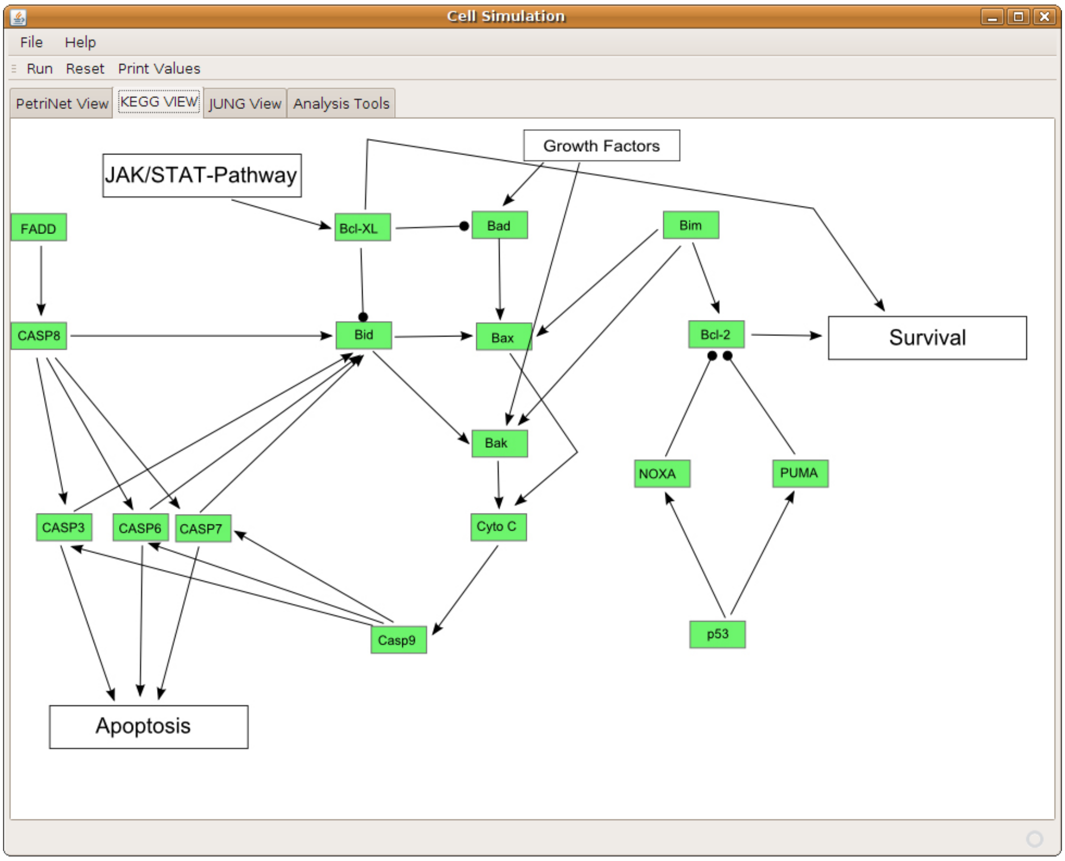


Figure 4.9: The most abstract view of Cell Sim. This view is heavily inspired by the KEGG and understandable by most users familiar with metabolite pathways.

Chapter 5

Discussion

At the beginning of this work the Tokyo declaration was cited which outlined the grand challenge of creating a virtual human. This virtual human is supposed to revolutionize the way drugs are developed leading to new cures against diseases. This goal motivated the three main objectives of this thesis: (1) the integration of methods and data, (2) the exploration of novel methods for the structural comparison of biomolecules, and (3) the evaluation of a novel concept that uses quantitative data to simulate interactions inside pathways based on a graph approach. This thesis presents several contributions to the main objectives.

SuperSite is the first publicly accessible data warehouse that integrates metabolite and drug data with extended information about proteins and their binding sites. It is valuable in evaluating characteristics of binding sites (conservation, comparative evaluation of apo and holo form) and for predicting a potential binding of drugs. It can be shown that SuperSite, e.g., successfully predicts a binding of the anti-cancer drug methotrexate at a dihydrofolate reductase. SuperSite could predict the comparative binding of the two ligands in the binding pocket, which is the method of action of this drug. The detection of binding partners, as exemplified with methotrexate, allows evaluation of potential desired and undesired effects of prospective drugs *in silico*. However, SuperSite as other tools presented in this thesis are currently optimized to be used by humans. A computerized virtual human must also be able to access data sources and collect information, needed to simulate perturbation experiments. For the realization of the virtual human to come true those approaches must be machine accessible via web services. Superimposé is a framework that integrates 3D molecular screening methods and related databases. To this end, three classes of similarity screening problems are defined that can be addressed. It can be shown that similarity screening methods and databases are suitable to answer questions of biological relevance. For instance, Superimposé is able to identify compounds with similar ATC classes in drug databases. This highlights that the 3D paradigm of Superimposé is able to retrieve useful compounds using a template compound. An advantage is that the framework is able to apply more than a single similarity screening problem. This can be important, as every method may have different advantages that can complement each other. Thus, the framework has a range of applications from drug development to the functional prediction of proteins. Superimposé is a central repository for methods and can also serve as independent benchmarking tool for various algorithms. The next versions of Superimposé will include an extended set of databases and methods. However, to cope with the computational demands, an extension of Superimposé to a cluster computing architecture becomes increasingly important.

LaJolla, an open source software framework, is able to structurally align a large set of biomolecules rapidly. This method also allows for the screening of proteins and RNAs based on structures and substructures. With the rapid growth of available macromolecular structures, fast screening becomes increasingly important. It can be shown that the method returns good results in comparison to state of the art methods in the field of structural protein alignment. The approach can be fast and accurate. It is able to outperform direct competitors in functional classification of RNA. Additionally, it solves all difficult RNA alignment examples discussed in the literature. However, whether the LaJolla core algorithm can be improved should be evaluated. It should be checked whether an enlargement of the initial seed n-gram alignment leads to better results. This is, for instance, applied by methods such as CE. Multi-processor architectures are becoming more widely used today, therefore the core algorithm of LaJolla must be parallelized. LaJolla should be used to answer questions in the field of structural biology that are currently neglected due to the fact that many methods available are not fast and accurate enough. It is, e.g., interesting to evaluate if the local secondary structure of small compound binding sites is more similar than the global structure of the same proteins. This is also interesting in the context of protein-protein docking. LaJolla can help to elucidate whether the notion of non-linear alignments exists in RNA structures; a formation known to exist in proteins. The RNA community lacks a (semi-) automatically generated structural classification, what is state of the art for proteins (among others CATH, SCOP). LaJolla is a suitable tool for generating and maintaining such a classification automatically. Finally, LaJolla should be made web accessible and connected to various databases. This can ideally be achieved by integrating LaJolla into the next version of Superimposé.

The last chapter of the thesis covers a method that uses Petri nets and RNA expression levels to model the signal flow inside a signal transduction network. The concept uses the human apoptosis pathway to predict whether a cell is undergoing cell death. Various options are evaluated that map expression levels in the network. The most promising concept is the usage of stochastic Petri nets with the firing probability of a transition representing the RNA expression level of a protein. It can be shown that this concept predicts a correct result regarding apoptosis when compared to the amount of pro- and anti-apoptotic proteins present in the cell. This highlights, on one hand, that cancer is not only dependent the anti apoptotic stimuli, on the other hand, that an extension to more pathways is needed for a comprehensive understanding of different cancer cell lines. To overcome the bias between transcription and translation, protein data should be used instead of the RNA expression level data. A novel simulation software, Cell Sim is presented that is able to carry out stochastic simulations of a Petri net. The software is able to hide the complexity of Petri nets by providing a presentation that can also be interpreted by non experts. To make the vision of a virtual human come true the network must be extended to more signaling networks and finally to a cell. To this end, the Petri net method, as well as Cell Sim, are important building blocks.

The Tokyo declaration mentions that the development of the virtual human will take “the next thirty years” (from 2008). This work – the integration of ligand characteristics for the simulation of cellular reactions – is only a single contribution in the initial phase of this grand challenge. The whole is more than the sum of its parts, as Aristotle outlined in Metaphysics.

Chapter 6

Appendix

Ehrenwörtliche Erklärung

Hiermit erkläre ich, dass ich diese Arbeit selbstständig verfasst und keine anderen als die angegebenen Hilfsmittel und Quellen verwendet habe.

Berlin, im Mai 2009 Raphael André Bauer

Zusammenfassung

Eine Hauptcharakteristik der Lebenswissenschaften ist die große Menge von Rohdaten, die von modernen biotechnologischen Verfahren erzeugt werden. Diese Rohdaten können beispielsweise von großen Experimentalreihen stammen, die untersuchen, ob bestimmte Moleküle gegen eine Krankheit wirken. Des Weiteren entstehen große Mengen von Rohdaten bei der Bestimmung der 3D Struktur von Biomolekülen und bei der Aufklärung des genetischen Codes von Lebewesen. Die Auswertung und Integration wird daher zunehmend wichtiger – zurzeit wird jedoch nur ein kleiner Teil dieser Rohdaten integriert.

Datenintegration ist Gegenstand des ersten Teils dieser Arbeit. Ein Data-Warehouse wird vorgestellt, das 3D Information von Proteinen mit Wissen über Medikamente und 3D Techniken zur Analyse bzw. Vorhersage von Bindungsstellen integriert. Die begrenzte Genauigkeit, mit der Moleküle in Datenbanken gesucht werden können, ist oftmals ein weiteres Problem. Aus diesem Grund wurde ein Framework entwickelt, das die Integration von Datenquellen und Methoden zur genauen 3D Suche ermöglicht.

Die Anzahl an 3D Strukturen von Makromolekülen wie Proteinen und RNS wächst rapide. Es gibt jedoch nur wenige Methoden, die eine 3D Ähnlichkeitssuche auf tausenden Proteinen in Echtzeit erlauben. Zudem gibt es nur wenige Methoden, die auf RNS Strukturen funktionieren. Diese Arbeit präsentiert eine neue Methode, die auf der Umwandlung von 3D Strukturen zu Zeichenketten basiert. Zur schnellen Suche wird eine Index Struktur verwendet. Es kann gezeigt werden, dass die Methode ähnliche oder bessere Resultate im Vergleich zu führenden Methoden im strukturellen Protein- und RNS-Alignment liefert. Die Methode kann in Hochdurchsatz-Experimenten verwendet werden, da Geschwindigkeit und Präzision eingestellt werden kann.

Der letzte Teil der Arbeit befasst sich mit der Interaktion in Signalwegen. RNS Expressionsraten von Proteinen korrelieren häufig mit der Anzahl von Signalen die von diesen Proteinen in einem biologischen Netzwerk generiert werden. Daher werden verschiedene Konzepte evaluiert, die die Expression von Genen auf den Signalweg der Apoptose übertragen. Die Abstraktion des Apoptose Signalwegs erfolgt durch ein Petri Netz. Zudem wird ein Softwarepaket vorgestellt, das Petri Netze simulieren kann. Die Software ist in der Lage, die Komplexität des Petri Netzes auszublenden, was ermöglicht, dass Nicht-Experten die Software benutzen.

Schlüsselwörter: Medikamentenwirkungen, Chemische Ähnlichkeit, Substruktursuche, Strukturelles Alignment, Protein, RNS, Hash-Tabellen, N-Gramme, Drehwinkel, Petri Netz, Microarray, Apoptose

Curriculum Vitae

For reasons of data protection, the curriculum vitae is not included in the online version.

List of Figures

List of Tables

Bibliography

[Abraham et al., 2008] Abraham, M., Dror, O., Nussinov, R., and Wolfson, H. J. J. (2008). Analysis and classification of RNA tertiary structures. *RNA*, 14(11):2274–2289.

[Alberts et al., 2002] Alberts, B., Johnson, A., Lewis, J., Raff, M., Roberts, K., and Walter, P. (2002). *Molecular Biology of the Cell, Fourth Edition*. Garland.

[Alon, 2006] Alon, U. (2006). *An Introduction to Systems Biology: Design Principles of Biological Circuits*. Chapman & Hall/CRC.

[Andreeva et al., 2008] Andreeva, A., Howorth, Chandonia, J. M., Brenner, S. E., Hubbard, T. J., Chothia, C., and Murzin, A. G. (2008). Data growth and its impact on the SCOP database: new developments. *Nucleic Acids Research*, 36(Database issue):D419–D425.

[Andreeva et al., 2004] Andreeva, A., Howorth, D., Brenner, S. E., Hubbard, T. J., Chothia, C., and Murzin, A. G. (2004). SCOP database in 2004: refinements integrate structure and sequence family data. *Nucleic Acids Research*, 32(Database issue):D226–D229.

[Aoki and Kanehisa, 2005] Aoki, K. F. and Kanehisa, M. (2005). Using the KEGG database resource. *Current Protocols in Bioinformatics*, Chapter 1(Unit 1.12).

[Barbosa and Horvath, 2004] Barbosa, F. and Horvath, D. (2004). Molecular similarity and property similarity. *Current Topics in Medicinal Chemistry*, 4(6):589–600.

[Barrett, 1996] Barrett, A. J. (1996). Nomenclature Committee of the International Union of Biochemistry and Molecular Biology (NC-IUBMB). Enzyme nomenclature. Recommendations 1992. Supplement 3: corrections and additions (1995). *European Journal of Biochemistry / FEBS*, 237(1):1–5.

[Bashton et al., 2007] Bashton, M., Nobeli, I., and Thornton, J. M. (2007). PROCOGNATE: a cognate ligand domain mapping for enzymes. *Nucleic Acids Research*, 36(Database issue):D618–D622.

[Bauer et al., 2008] Bauer, R. A., Rother, K., Bujncki, J., and Preissner, R. (2008). Suffix techniques as a rapid method for RNA substructure search. *Genome Informatics*, 20:183–198.

[Bayry et al., 2008] Bayry, J., Tchilian, E. Z., Davies, M. N., Forbes, E. K., Draper, S. J., Kaveri, S. V., Hill, A. V., Kazatchkine, M. D., Beverley, P. C., Flower, D. R., and Tough, D. F. (2008). In silico identified CCR4 antagonists target regulatory T cells and exert adjuvant activity in vaccination. *Proceedings of the National Academy of Sciences*, 105(29):10221–10226.

[Bekaert et al., 2003] Bekaert, M., Bidou, L., Denise, A., Duchateau-Nguyen, G., Forest, J. P., Froidevaux, C., Hatin, I., Rousset, J. P., and Termier, M. (2003). Towards a computational model for -1 eukaryotic frameshifting sites. *Bioinformatics*, 19(3):327–335.

[Benson et al., 2008] Benson, M. L., Smith, R. D., Khazanov, N. A., Dimcheff, B., Beaver, J., Dresslar, P., Nerothin, J., and Carlson, H. A. (2008). Binding MOAD, a high-quality protein-ligand database. *Nucleic Acids Research*, 36(Database issue):D674–D678.

[Berkman et al., 2008] Berkman, M. B., Pacheco, J. S., and Plutzer, E. (2008). Evolution and Creationism in America's Classrooms: A National Portrait. *PLoS Biology*, 6(5):e124+.

[Berman et al., 2007] Berman, H., Henrick, K., Nakamura, H., and Markley, J. L. (2007). The worldwide Protein Data Bank (wwPDB): ensuring a single, uniform archive of PDB data. *Nucleic Acids Research*, 35(Database issue):D301–D303.

[Berman et al., 2000] Berman, H. M., Westbrook, J., Feng, Z., Gilliland, G., Bhat, T. N., Weissig, H., Shindyalov, I. N., and Bourne, P. E. (2000). The Protein Data Bank. *Nucleic Acids Research*, 28(1):235–242.

[Berman et al., 2002] Berman, H. M., Westbrook, J., Feng, Z., Iype, L., Schneider, B., and Zardecki, C. (2002). The Nucleic Acid Database. *Acta Crystallographica Section D*, 58(6 Part 1):889–898.

[Bonetta, 2008] Bonetta, L. (2008). Epigenomics: Detailed analysis. *Nature*, 454(7205):795–798.

[Borden et al., 2008] Borden, E. C., Kluger, H., and Crowley, J. (2008). Apoptosis: a clinical perspective. *Nature Reviews Drug Discovery*, 7(12):959.

[Bowler, 2009] Bowler, P. J. (2009). Darwin's originality. *Science*, 323(5911):223–226.

[Burkhardt et al., 1999] Burkhardt, S., Crauser, A., Ferragina, P., Lenhof, H. P., Rivals, E., and Vingron, M. (1999). q-gram based database searching using a suffix array (QUASAR). In *RECOMB '99: Proceedings of the third annual international conference on Computational molecular biology*, pages 77–83, New York, NY, USA. ACM.

[Butcher, 2005] Butcher, E. C. (2005). Innovation: Can cell systems biology rescue drug discovery? *Nature Reviews Drug Discovery*, 4(6):461–467.

[Capriotti and Marti-Renom, 2008] Capriotti, E. and Marti-Renom, M. A. (2008). RNA structure alignment by a unit-vector approach. *Bioinformatics*, 24(16):112–118.

[Chakrabarti and Lanczycki, 2007] Chakrabarti, S. and Lanczycki, C. J. (2007). Analysis and prediction of functionally important sites in proteins. *Protein Science*, 16(1):4–13.

[Chandonia et al., 2004] Chandonia, J. M., Hon, G., Walker, N. S., Lo Conte, L., Koehl, P., Levitt, M., and Brenner, S. E. (2004). The astral compendium in 2004. *Nucleic Acids Research*, 32(Database issue):D189–D192.

[Chang et al., 2008] Chang, Y.-F. F., Huang, Y.-L. L., and Chin (2008). SARSA: a web tool for structural alignment of RNA using a structural alphabet. *Nucleic Acids Research*, 36(Web Server issue):19–24.

[Chaouiya, 2007] Chaouiya, C. (2007). Petri net modelling of biological networks. *Briefings in Bioinformatics*, 8(4):210–219.

[Cheek et al., 2004] Cheek, S., Qi, Y., Krishna, S. S., Kinch, L. N., and Grishin, N. V. (2004). SCOPmap: automated assignment of protein structures to evolutionary superfamilies. *BMC Bioinformatics*, 5(1):197+.

[Chen et al., 2001] Chen, X., Liu, M., and Gilson, M. K. (2001). BindingDB: a web-accessible molecular recognition database. *Combinatorial Chemistry & High Throughput Screening*, 4(8):719–725.

[Chen and Reynolds, 2002] Chen, X. and Reynolds, C. H. (2002). Performance of similarity measures in 2D fragment-based similarity searching: comparison of structural descriptors and similarity coefficients. *Journal of Chemical Information and Computer Sciences*, 42(6):1407–1414.

[Chu and Chen, 2008] Chu, L. H. and Chen, B. S. (2008). Construction of a cancer-perturbed protein-protein interaction network for discovery of apoptosis drug targets. *BMC Systems Biology*, 2(1):56+.

[Cohen et al., 2008] Cohen, A. A., Geva-Zatorsky, N., Eden, E., Frenkel-Morgenstern, M., Issaeva, I., Sigal, A., Milo, R., Cohen-Saidon, C., Liron, Y., Kam, Z., Cohen, L., Danon, T., Perzov, N., and Alon, U. (2008). Dynamic proteomics of individual cancer cells in response to a drug. *Science*, 322(5907):1516+.

[Cormen et al., 2003] Cormen, T. H., Leiserson, C. E., Rivest, R. L., and Stein, C. (2003). *Introduction to Algorithms*. McGraw-Hill Science / Engineering / Math, 2nd edition.

[Crespo et al., 2008] Crespo, A., Zhang, X., and Fernández, A. (2008). Redesigning Kinase Inhibitors to Enhance Specificity. *Journal of Medicinal Chemistry*, 51(16):4890–4898.

[Cuff et al., 2009] Cuff, A. L., Sillitoe, I., Lewis, T., Redfern, O. C., Garratt, R., Thornton, J., and Orengo, C. A. (2009). The CATH classification revisited–architectures reviewed and new ways to characterize structural divergence in superfamilies. *Nucleic Acids Research*, 37(Database issue):D310–D314.

[Davidov et al., 2003] Davidov, E., Holland, J., Marple, E., and Naylor, S. (2003). Advancing drug discovery through systems biology. *Drug Discovery Today*, 8(4):175–183.

[Dietzfelbinger et al., 1988] Dietzfelbinger, M., Karlin, A. R., Mehlhorn, K., Friedhelm, Rohnert, H., and Tarjan, R. E. (1988). Dynamic perfect hashing: Upper and lower bounds. In *IEEE Symposium on Foundations of Computer Science*, pages 524–531.

[Dodge et al., 1998] Dodge, C., Schneider, R., and Sander, C. (1998). The HSSP database of protein structure-sequence alignments and family profiles. *Nucleic Acids Research*, 26(1):313–315.

[Dowell and Eddy, 2006] Dowell, R. D. and Eddy, S. R. (2006). Efficient pairwise RNA structure prediction and alignment using sequence alignment constraints. *BMC Bioinformatics*, 7:400+.

[Dror et al., 2006] Dror, O., Nussinov, R., and Wolfson, H. J. (2006). The ARTS web server for aligning RNA tertiary structures. *Nucleic Acids Research*, 34(Web Server issue).

[Duarte and Pyle, 1998] Duarte, C. M. and Pyle, A. M. (1998). Stepping through an RNA structure: a novel approach to conformational analysis. *Journal of Molecular Biology*, 284(5):1465–1478.

[Dunkel et al., 2008] Dunkel, M., Günther, S., Ahmed, J., Wittig, B., and Preissner, R. (2008). SuperPred: drug classification and target prediction. *Nucleic Acids Research*, 36(Web Server issue).

[Eddy, 2001] Eddy, S. R. (2001). Non-coding RNA genes and the modern RNA world. *Nature Reviews Genetics*, 2(12):919–929.

[Editorial, 2009] Editorial (2009). Data for the masses. *Nature*, 457(7226):129.

[Elmore, 2007] Elmore, S. (2007). Apoptosis: a review of programmed cell death. *Toxicol Pathol*, 35(4):495–516.

[Ender et al., 2008] Ender, C., Krek, A., Friedländer, M. R., Beitzinger, M., Weinmann, L., Chen, W., Pfeffer, S., Rajewsky, N., and Meister, G. (2008). A Human snoRNA with MicroRNA-Like Functions. *Molecular Cell*, 32(4):519–528.

[Feng et al., 2004] Feng, Z., Chen, L., Maddula, H., Akcan, O., Oughtred, R., Berman, H. M., and Westbrook, J. (2004). Ligand Depot: a data warehouse for ligands bound to macromolecules. *Bioinformatics*, 20(13):2153–2155.

[Ferrè et al., 2007] Ferrè, F., Ponty, Y., Lorenz, W. A., and Clote, P. (2007). DIAL: a web server for the pairwise alignment of two RNA three-dimensional structures using nucleotide, dihedral angle and base-pairing similarities. *Nucleic Acids Research*, 35(Web Server issue):W659–W668.

[Flemming, 2008] Flemming, A. (2008). Apoptosis: New strategies to tip the BCL-2 balance. *Nature Reviews Drug Discovery*, 7(12):977.

[Formella, 2005] Formella, A. (2005). Approximate point set match for partial protein structure alignment. In *Proceedings of Bioinformatics: Knowledge Discovery in Biology (BKDB2005)*, pages 53–57.

[Friedberg et al., 2007] Friedberg, I., Harder, T., Sitbon, E., Li, Z., and Godzik, A. (2007). Using an alignment of fragment strings for comparing protein structures. *Bioinformatics*, 23(2):219–224.

[Frömmel et al., 2003] Frömmel, C., Gille, C., Goede, A., Gropl, C., Hougardy, S., Nierhoff, T., Preissner, R., and Thimm, M. (2003). Accelerating screening of 3D protein data with a graph theoretical approach. *Bioinformatics*, 19(18):2442–2447.

[Gao and Zaki, 2008] Gao, F. and Zaki, M. J. (2008). PSIST: A scalable approach to indexing protein structures using suffix trees. *Journal of Parallel and Distributed Computing*, 68(1):54–63.

[García and Rodríguez, 2002] García and Rodríguez (2002). New sequential and parallel derivative–free algorithms for unconstraint optimization. *SIAM Journal on Optimization*, 13(1):79–96.

[Giegé, 2008] Giegé, R. (2008). Toward a more complete view of tRNA biology. *Nature Structural & Molecular Biology*, 15(10):1007–1014.

[Giegerich and Kurtz, 1997] Giegerich, R. and Kurtz, S. (1997). From Ukkonen to McCreight and Weiner: A Unifying View of Linear-Time Suffix Tree Construction. *Algorithmica*, 19(3):331–353.

[Gille and Frömmel, 2001] Gille, C. and Frömmel, C. (2001). STRAP: editor for STRuctural Alignments of Proteins. *Bioinformatics*, 17(4):377–378.

[Goede et al., 2005a] Goede, A., Dunkel, M., Frömmel, C., and Preissner, R. (2005a). SuperDrug: a conformational drug database. *Bioinformatics*, 21(9):1751–1753.

[Goede et al., 2005b] Goede, A., Jaeger, I. S., and Preissner, R. (2005b). Superficial–surface mapping of proteins via structure-based peptide library design. *BMC Bioinformatics*, 6:223.

[Gold and Jackson, 2006] Gold, N. D. and Jackson, R. M. (2006). SitesBase: a database for structure-based protein-ligand binding site comparisons. *Nucleic Acids Research*, 34(Database issue):231–234.

[Grafahrend-Belau et al., 2008] Grafahrend-Belau, E., Schreiber, F., Heiner, M., Sackmann, A., Junker, B., Grunwald, S., Speer, A., Winder, K., and Koch, I. (2008). Modularization of biochemical networks based on classification of Petri net t-invariants. *BMC Bioinformatics*, 9(1):90+.

[Green and Chamberlain, 2009] Green, M. and Chamberlain, M. (2009). Renal dysfunction during and after high-dose methotrexate. *Cancer Chemotherapy and Pharmacology*, 63(4):599–604.

[Greene et al., 2007] Greene, L. H., Lewis, T. E., Addou, S., Cuff, A., Dallman, T., Dibley, M., Redfern, O., Pearl, F., Nambudiry, R., Reid, A., Sillitoe, I., Yeats, C., Thornton, J. M., and Orengo, C. A. (2007). The CATH domain structure database: new protocols and classification levels give a more comprehensive resource for exploring evolution. *Nucleic Acids Research*, 35(Database issue):D291–D297.

[Grunwald et al., 2008] Grunwald, S., Speer, A., Ackermann, J., and Koch, I. (2008). Petri net modelling of gene regulation of the Duchenne muscular dystrophy. *Biosystems*, 92(2):189–205.

[Guerler and Knapp, 2008] Guerler, A. and Knapp, E.-W. (2008). Novel protein folds and their nonsequential structural analogs. *Protein Science*, 17(8):1374–1382.

[Guha et al., 2006] Guha, R., Howard, M. T., Hutchison, G. R., Rust, M. P., Rzepa, H., Steinbeck, C., Wegner, J. K., and Willighagen, E. L. (2006). The Blue Obelisk–Interoperability in Chemical Informatics. *Journal of Chemical Information and Modeling*, 46:991–998.

[Günther et al., 2007] Günther, S., May, P., Hoppe, A., Frömmel, C., and Preissner, R. (2007). Docking without docking: ISEARCH-prediction of interactions using known interfaces. *Proteins*, 69(4):839–844.

[Gusfield, 1997] Gusfield, D. (1997). *Algorithms on Strings, Trees, and Sequences: Computer Science and Computational Biology.* Cambridge University Press.

[Guyon et al., 2004] Guyon, F., Camproux, A. C., Hochez, J., and Tufféry, P. (2004). SA-Search: a web tool for protein structure mining based on a Structural Alphabet. *Nucleic Acids Research*, 32(Web Server issue):W545–W548.

[He et al., 2008] He, Y., Chen, Y., Alexander, P., Bryan, P. N., and Orban, J. (2008). NMR structures of two designed proteins with high sequence identity but different fold and function. *Proceedings of the National Academy of Sciences*, 105(38):14412–14417.

[Heiner et al., 2004] Heiner, M., Koch, I., and Will, J. (2004). Model validation of biological pathways using Petri nets - demonstrated for apoptosis. *Biosystems*, 75(1-3):15–28.

[Heiner et al., 2008] Heiner, M., Richter, R., Rohr, C., and Schwarick, M. (2008). Snoopy - a tool to design and execute graph-based formalisms. *Petri Net Newsletter*, 74:8–22.

[Hodis et al., 2008] Hodis, E., Prilusky, J., Martz, E., Silman, I., Moult, J., and Sussman, J. L. (2008). Proteopedia - a scientific wiki bridging the rift between 3D structure and function of biomacromolecules. *Genome Biology*, 9:R121+.

[Hofacker, 2003] Hofacker, I. L. (2003). Vienna RNA secondary structure server. *Nucleic Acids Research*, 31(13):3429–3431.

[Hofacker et al., 2004] Hofacker, I. L., Bernhart, S. H., and Stadler, P. F. (2004). Alignment of RNA base pairing probability matrices. *Bioinformatics*, 20(14):2222–2227.

[Hoppe and Frömmel, 2003] Hoppe, A. and Frömmel, C. (2003). Needlehaystack: A program for the rapid recognition of local structures in large sets of atomic coordinates. *Journal of Applied Crystallography*, 36:1090–1097.

[Huang and Schroeder, 2006] Huang, B. and Schroeder, M. (2006). LIGSITEcsc: Predicting ligand binding sites using the Connolly surface and degree of conservation. *BMC Structural Biology*, 6:19+.

[Huang and Strasser, 2000] Huang, D. and Strasser, A. (2000). BH3-Only Proteins: Essential Initiators of Apoptotic Cell Death. *Cell*, 103(6):839–842.

[Ihlenfeldt et al., 2002] Ihlenfeldt, W. D., Voigt, J. H., Bienfait, B., Oellien, F., and Nicklaus, M. C. (2002). Enhanced CACTVS browser of the Open NCI Database. *Journal of Chemical Information and Computer Sciences*, 42(1):46–57.

[Ilyin et al., 2004] Ilyin, V. A., Abyzov, A., and Leslin, C. M. (2004). Structural alignment of proteins by a novel TOPOFIT method, as a superimposition of common volumes at a topomax point. *Protein Science*, 13(7):1865–1874.

[Kabsch, 1976] Kabsch, W. (1976). A solution for the best rotation to relate two sets of vectors. *Acta Crystallographica Section A*, 32(5):922–923.

[Kanehisa et al., 2007] Kanehisa, M., Araki, M., Goto, S., Hattori, M., Hirakawa, M., Itoh, M., Katayama, T., Kawashima, S., Okuda, S., Tokimatsu, T., and Yamanishi, Y. (2007). KEGG for linking genomes to life and the environment. *Nucleic Acids Research*, 36(Database issue):D480–D484.

[Kaye, 2008] Kaye, J. (2008). The regulation of direct-to-consumer genetic tests. *Human molecular genetics*, 17(R2):R180–R183.

[Kell, 2007] Kell, D. B. (2007). The virtual human: Towards a global systems biology of multiscale, distributed biochemical network models. *IUBMB Life*, 59(11):689–695.

[Keller et al., 2008] Keller, A., Backes, C., Awadhi, M. A., Gerasch, A., Kuntzer, J., Kohlbacher, O., Kaufmann, M., and Lenhof, H. P. (2008). GeneTrailExpress: a web-based pipeline for the statistical evaluation of microarray experiments. *BMC Bioinformatics*, 9(1):552+.

[Kendrew et al., 1958] Kendrew, J. C., Bodo, G., Dintzis, H. M., Parrish, R. G., Wyckoff, H., and Phillips, D. C. (1958). A three-dimensional model of the myoglobin molecule obtained by x-ray analysis. *Nature*, 181(4610):662–666.

[Kirchner, 2007] Kirchner, S. (2007). An FPTAS for Computing the Similarity of three-dimensional Point Sets. *International Journal of Computational Geometry and Applications*, 17(2):161–174.

[Kitano, 2001] Kitano, H. (2001). *Foundations of Systems Biology.* The MIT Press.

[Kitano, 2002] Kitano, H. (2002). Systems biology: a brief overview. *Science*, 295(5560):1662–1664.

[Kolbeck et al., 2006] Kolbeck, B., May, P., Schmidt-Goenner, T., Steinke, T., and Knapp, E.-W. (2006). Connectivity independent protein fold detection: a hierarchical approach. *BMC Bioinformatics*, 7:510+.

[Kolodny et al., 2005] Kolodny, R., Koehl, P., and Levitt, M. (2005). Comprehensive evaluation of protein structure alignment methods: scoring by geometric measures. *Journal of Molecular Biology*, 346(4):1173–1188.

[Krissinel and Henrick, 2004] Krissinel, E. and Henrick, K. (2004). Secondary-structure matching (SSM), a new tool for fast protein structure alignment in three dimensions. *Acta Crystallographica Section D. Biological Crystallography*, 60(Pt 12 Pt 1):2256–2268.

[Kubinyi, 1998a] Kubinyi, H. (1998a). Molecular similarity. 1. Chemical structure and biological action. *Pharmazie Unserer Zeit*, 27(3):92–106.

[Kubinyi, 1998b] Kubinyi, H. (1998b). Molecular similarity. 2. The structural basis of drug design. *Pharmazie Unserer Zeit*, 27(4):158–172.

[Kume et al., 1991] Kume, A., Koyata, H., Sakakibara, T., Ishiguro, Y., Kure, S., and Hiraga, K. (1991). The glycine cleavage system. Molecular cloning of the chicken and human glycine decarboxylase cDNAs and some characteristics involved in the deduced protein structures. *Journal of Biological Chemistry*, 266(5):3323–3329.

[Laederach, 2007] Laederach, A. (2007). Informatics challenges in structured RNA. *Briefings in Bioinformatics*, 8(5):294–303.

[Laskowski, 2001] Laskowski, R. A. (2001). PDBsum: summaries and analyses of PDB structures. *Nucleic Acids Research*, 29(1):221–222.

[Laskowski, 2007] Laskowski, R. A. (2007). Enhancing the functional annotation of pdb structures in pdbsum using key figures extracted from the literature. *Bioinformatics*, 23(14):1824–1827.

[Laskowski et al., 2005] Laskowski, R. A., Chistyakov, V. V., and Thornton, J. M. (2005). PDBsum more: new summaries and analyses of the known 3D structures of proteins and nucleic acids. *Nucleic Acids Research*, 33(Database issue):D266+.

[Lathrop, 1994] Lathrop, R. H. (1994). The protein threading problem with sequence amino acid interaction preferences is NP-complete. *Protein Engineering*, 7(9):1059–1068.

[Lee et al., 2004] Lee, D. Y., Zimmer, R., Lee, S. Y., Hanisch, D., and Park, S. (2004). Knowledge representation model for systems-level analysis of signal transduction networks. *Genome Informatics*, 15(2):234–243.

[Lee et al., 2006] Lee, D.-Y., Zimmer, R., Lee, S. Y., and Park, S. (2006). Colored petri net modeling and simulation of signal transduction pathways. *Metabolic Engineering*, 8(2):112–122.

[Lehninger et al., 2008] Lehninger, A., Nelson, D. L., and Cox, M. M. (2008). *Principles of Biochemistry.* W. H. Freeman, fifth edition.

[Leontis et al., 2006] Leontis, N. B., Altman, R. B., Berman, H. M., Brenner, S. E., Brown, J. W., Engelke, D. R., Harvey, S. C., Holbrook, S. R., Jossinet, F., Lewis, S. E., Major, F., Mathews, D. H., Richardson, J. S., Williamson, J. R., and Westhof, E. (2006). The RNA Ontology Consortium: an open invitation to the RNA community. *RNA*, 12(4):533–541.

[Lescoute et al., 2005] Lescoute, A., Leontis, N. B., Massire, C., and Westhof, E. (2005). Recurrent structural RNA motifs, Isostericity Matrices and sequence alignments. *Nucleic Acids Research*, 33(8):2395–2409.

[Lescoute and Westhof, 2006] Lescoute, A. and Westhof, E. (2006). The interaction networks of structured RNAs. *Nucleic Acids Research*, 34(22):6587–6604.

[Leslin et al., 2007] Leslin, C. M., Abyzov, A., and Ilyin, V. A. (2007). TOPOFIT-DB, a database of protein structural alignments based on the TOPOFIT method. *Nucleic Acids Research*, 35(Database issue):D317–D321.

[Levitt, 2007] Levitt, M. (2007). Growth of novel protein structural data. *Proceedings of the National Academy of Sciences*, 104(9):3183–3188.

[Li et al., 2006] Li, Weizhong, Godzik, and Adam (2006). Cd-hit: a fast program for clustering and comparing large sets of protein or nucleotide sequences. *Bioinformatics*, 22(13):1658–1659.

[Lim et al., 2003] Lim, L. P., Glasner, M. E., Yekta, S., Burge, C. B., and Bartel, D. P. (2003). Vertebrate MicroRNA Genes. *Science*, 299(5612):1540+.

[Lipinski et al., 2001] Lipinski, C. A., Lombardo, F., Dominy, B. W., and Feeney, P. J. (2001). Experimental and computational approaches to estimate solubility and permeability in drug discovery and development settings. *Advanced Drug Delivery Reviews*, 46(1-3):3–26.

[Lo et al., 2007] Lo, W.-C., Huang, P.-J., Chang, C.-H., and Lyu, P.-C. (2007). Protein structural similarity search by Ramachandran codes. *BMC Bioinformatics*, 8:307+.

[Lo Conte et al., 2002] Lo Conte, L., Brenner, S. E., Hubbard, T. J. P., Chothia, C., and Murzin, A. G. (2002). SCOP database in 2002: refinements accommodate structural genomics. *Nucleic Acids Research*, 30(Database issue):D264–D267.

[Locasale et al., 2009] Locasale, J., Napoli, A., Chen, S., Berman, H., and Lawson, C. (2009). Signatures of protein-DNA recognition in free DNA binding sites. *Journal of Molecular Biology*, 386(4):1054–1065.

[Maiti et al., 2004] Maiti, R., Van Domselaar, G. H., Zhang, H., and Wishart, D. S. (2004). Superpose: a simple server for sophisticated structural superposition. *Nucleic Acids Research*, 32(Web Server issue):W590–W594.

[Mardis, 2008] Mardis, E. R. R. (2008). Next-Generation DNA Sequencing Methods. *Annual Review of Genomics and Human Genetics*, 9:387–402.

[Martin et al., 2002] Martin, Y. C., Kofron, J. L., and Traphagen, L. M. (2002). Do structurally similar molecules have similar biological activity? *Journal of Medicinal Chemistry*, 45(19):4350–4358.

[Medzhitov and Littman, 2008] Medzhitov, R. and Littman, D. (2008). HIV immunology needs a new direction. *Nature*, 455(7213):591.

[Minai et al., 2008] Minai, R., Matsuo, Y., Onuki, H., and Hirota, H. (2008). Method for comparing the structures of protein ligand-binding sites and application for predicting protein-drug interactions. *Proteins*, 72(1):367–381.

[Morgan, 2001] Morgan, M. J. (2001). Initial sequencing and analysis of the human genome. *Nature*, 409(6822):860–921.

[Murray et al., 2005] Murray, L. J. W., Richardson, J. S., Iii, A. W. B., and Richardson, D. C. (2005). RNA backbone rotamers – finding your way in seven dimensions. *Biochemical Society Transactions*, pages 485–487.

[News, 2008] News (2008). Systems biologists hatch plan for virtual human. *Nature*, 451(7181):879+.

[Nissen et al., 2000] Nissen, P., Hansen, J., Ban, N., Moore, P. B., and Steitz, T. A. (2000). The structural basis of ribosome activity in peptide bond synthesis. *Science*, 289(5481):920–930.

[Noble, 2002] Noble, D. (2002). Modeling the heart–from genes to cells to the whole organ. *Science*, 295(5560):1678–1682.

[Novotny et al., 2004] Novotny, M., Madsen, D., and Kleywegt, G. J. (2004). Evaluation of protein fold comparison servers. *Proteins*, 54(2):260–270.

[Okuda et al., 2008] Okuda, S., Yamada, T., Hamajima, M., Itoh, M., Katayama, T., Bork, P., Goto, S., and Kanehisa, M. (2008). KEGG Atlas mapping for global analysis of metabolic pathways. *Nucleic Acids Research*, 36(Web Server issue):W423–W426.

[Pandit and Skolnick, 2008] Pandit, S. B. and Skolnick, J. (2008). Fr-TM-align: A new protein structural alignment method based on fragment alignments and the TM-score. *BMC Bioinformatics*, 9:531+.

[Papin et al., 2005] Papin, J. A., Hunter, T., Palsson, B. O., and Subramaniam, S. (2005). Reconstruction of cellular signalling networks and analysis of their properties. *Nature Reviews Molecular Cell Biology*, 6(2):99–111.

[Parisien and Major, 2008] Parisien, M. and Major, F. (2008). The MC-Fold and MC-Sym pipeline infers RNA structure from sequence data. *Nature*, 452(7183):51–55.

[Park et al., 2008] Park, J. H., Scheerer, P., Hofmann, K. P., Choe, H.-W., and Ernst, O. P. (2008). Crystal structure of the ligand-free G-protein-coupled receptor opsin. *Nature*, 454(7201):183–187.

[Peleg et al., 2005] Peleg, M., Rubin, D., and Altman, R. B. (2005). Using Petri Net tools to study properties and dynamics of biological systems. *Journal of the American Medical Informatics Association*, 12(2):181–199.

[Penchovsky and Breaker, 2005] Penchovsky, R. and Breaker, R. R. (2005). Computational design and experimental validation of oligonucleotide-sensing allosteric ribozymes. *Nature Biotechnology*, 23(11):1424–1433.

[Pentchev, 2008] Pentchev, K. (2008). Graph-based Simulation of Apoptosis using Petri Nets. *Free University Berlin Bachelor's Thesis*.

[Preissner et al., 1999] Preissner, R., Goede, A., and Frommel, C. (1999). Homonyms and synonyms in the dictionary of interfaces in proteins (dip). *Bioinformatics*, 15(10):832–836.

[Preissner et al., 2001] Preissner, R., Goede, A., Rother, K., Osterkamp, F., Koert, U., and Froemmel, C. (2001). Matching organic libraries with protein-substructures. *Journal of Computer-Aided Molecular Design*, 15(9):811–817.

[Ramachandran et al., 1963] Ramachandran, G. N., Ramakrishnan, C., and Sasisekharan, V. (1963). Stereochemistry of polypeptide chain configurations. *Journal of Molecular Biology*, 7:95–99.

[Rao and Rossmann, 1973] Rao, S. T. and Rossmann, M. G. (1973). Comparison of super-secondary structures in proteins. *Journal of Molecular Biology*, 76(2):241–256.

[Rasmussen et al., 2007] Rasmussen, S. G. F., Choi, H.-J., Rosenbaum, D. M., Kobilka, T. S., Thian, F. S., Edwards, P. C., Burghammer, M., Ratnala, V. R. P., Sanishvili, R., Fischetti, R. F., Schertler, G. F. X., Weis, W. I., and Kobilka, B. K. (2007). Crystal structure of the human beta 2 adrenergic G-protein-coupled receptor. *Nature*, 450(7168):383–387.

[Reeder et al., 2006] Reeder, J., Höchsmann, M., Rehmsmeier, M., Voss, B., and Giegerich, R. (2006). Beyond Mfold: Recent advances in RNA bioinformatics. *Journal of Biotechnology*, 124(1):41–55.

[Reichert and Suhnel, 2002] Reichert, J. and Suhnel, J. (2002). The IMB Jena Image Library of Biological Macromolecules: 2002 update. *Nucleic Acids Research*, 30(1):253–254.

[Reisig, 1985] Reisig, W. (1985). *Petri Nets. An Introduction*, volume 4. Springer-Verlag GmbH.

[Richardson et al., 2008] Richardson, J. S., Schneider, B., Murray, L. W., Kapral, G. J., Immormino, R. M., Headd, J. J., Richardson, D. C., Ham, D., Hershkovits, E., Williams, L. D., Keating, K. S., Pyle, A. M., Micallef, D., Westbrook, J., and Berman, H. M. (2008). RNA backbone: consensus all-angle conformers and modular string nomenclature (an RNA Ontology Consortium contribution). *RNA*, 14(3):465–481.

[Ross et al., 2000] Ross, D. T., Scherf, U., Eisen, M. B., Perou, C. M., Rees, C., Spellman, P., Iyer, V., Jeffrey, S. S., Van de Rijn, M., Waltham, M., Pergamenschikov, A., Lee, J. C., Lashkari, D., Shalon, D., Myers, T. G., Weinstein, J. N., Botstein, D., and Brown, P. O. (2000). Systematic variation in gene expression patterns in human cancer cell lines. *Nature Genetics*, 24(3):227–235.

[Rother et al., 2005] Rother, K., Michalsky, E., and Leser, U. (2005). How well are protein structures annotated in secondary databases? *Proteins*, 60(4):571–576.

[Ruths et al., 2008] Ruths, D., Muller, M., Tseng, J. T., Nakhleh, L., and Ram, P. T. (2008). The signaling Petri net-based simulator: a non-parametric strategy for characterizing the dynamics of cell-specific signaling networks. *PLoS Computational Biology*, 4(2).

[Sackmann et al., 2006] Sackmann, A., Heiner, M., and Koch, I. (2006). Application of Petri net based analysis techniques to signal transduction pathways. *BMC Bioinformatics*, 7:482+.

[Sanger et al., 1977] Sanger, F., Nicklen, S., and Coulson, A. R. (1977). DNA sequencing with chain-terminating inhibitors. *Proceedings of the National Academy of Sciences*, 74(12):5463–5467.

[Scheerer et al., 2008] Scheerer, P., Park, J. H., Hildebrand, P. W., Kim, Y. J., Krausz, N., Choe, H.-W., Hofmann, K. P., and Ernst, O. P. (2008). Crystal structure of opsin in its G-protein-interacting conformation. *Nature*, 455(7212):497–502.

[Scheiber et al., 2009] Scheiber, J., Chen, B., Milik, M., Sukuru, S. C., Bender, A., Mikhailov, D., Whitebread, S., Hamon, J., Azzaoui, K., Urban, L., Glick, M., Davies, J. W., and Jenkins, J. L. (2009). Gaining Insight into Off-Target Mediated Effects of Drug Candidates with a Comprehensive Systems Chemical Biology Analysis. *Journal of Chemical Information and Modeling*, 49(2):308–317.

[Scholz, 2008] Scholz, C. (2008). Generierung eines Petri-Netzes zur quantitativen Simulation von Signalkaskaden der Apoptose. *Free University Berlin Master's Thesis*.

[Schormann et al., 2008] Schormann, N., Senkovich, O., Walker, K., Wright, D. L., Anderson, A. C., Rosowsky, A., Ananthan, S., Shinkre, B., Velu, S., and Chattopadhyay, D. (2008). Structure-based approach to pharmacophore identification, in silico screening, and three-dimensional quantitative structure-activity relationship studies for inhibitors of Trypanosoma cruzi dihydrofolate reductase function. *Proteins*, 73(4):889–901.

[Service, 2008] Service, R. F. (2008). Structural biology. Protein structure initiative: phase 3 or phase out. *Science*, 319(5870):1610–1613.

[Shakhnovich, 2006] Shakhnovich, E. (2006). Protein folding thermodynamics and dynamics: Where physics, chemistry, and biology meet. *Chemical Reviews*, 106(5):1559–1588.

[Shindyalov and Bourne, 1998] Shindyalov, I. N. and Bourne, P. E. (1998). Protein structure alignment by incremental combinatorial extension (CE) of the optimal path. *Protein Engeering*, 11(9):739–747.

[Shulman-Peleg et al., 2008] Shulman-Peleg, A., Shatsky, M., Nussinov, R., and Wolfson, H. J. J. (2008). MultiBind and MAPPIS: webservers for multiple alignment of protein 3D-binding sites and their interactions. *Nucleic Acids Research*, 36(Web Server issue):W260–W264.

[Sippl and Wiederstein, 2008] Sippl, M. J. and Wiederstein, M. (2008). A note on difficult structure alignment problems. *Bioinformatics*, 24(3):426–427.

[Stein, 2008] Stein, L. D. (2008). Bioinformatics: alive and kicking. *Genome Biology*, 9:114+.

[Stombaugh et al., 2009] Stombaugh, J., Zirbel, C. L., Westhof, E., and Leontis, N. B. (2009). Frequency and isostericity of RNA base pairs. *Nucleic Acids Research*, 37(7):2294–2312.

[Storey, 2008] Storey, S. (2008). Targeting apoptosis: selected anticancer strategies. *Nature Reviews Drug Discovery*, 7(12):971–972.

[Storz, 2002] Storz, G. (2002). An expanding universe of noncoding RNAs. *Science*, 296(5571):1260–1263.

[Sumathi et al., 2006] Sumathi, K., Ananthalakshmi, P., Md, and Sekar, K. (2006). 3dSS: 3D structural superposition. *Nucleic Acids Research*, 34(Web Server issue):W128–W132.

[Tamura et al., 2004] Tamura, M., Hendrix, D. K., Klosterman, P. S., Schimmelman, N. R., Brenner, S. E., and Holbrook, S. R. (2004). SCOR: Structural Classification of RNA, version 2.0. *Nucleic Acids Research*, 32(Database issue):D182–D184.

[Täubig et al., 2006] Täubig, H., Buchner, A., and Griebsch, J. (2006). PAST: fast structure-based searching in the PDB. *Nucleic Acids Research*, 34(Web Server issue):W20–W23.

[Thimm et al., 2004] Thimm, M., Goede, A., Hougardy, S., and Preissner, R. (2004). Comparison of 2D similarity and 3D superposition. Application to searching a conformational drug database. *Journal of Chemical Information and Modeling*, 44(5):1816–1822.

[Tikhonova et al., 2008] Tikhonova, I. G., Sum, C. S., Neumann, S., Engel, S., Raaka, B. M., Costanzi, S., and Gershengorn, M. C. (2008). Discovery of novel agonists and antagonists of the free fatty acid receptor 1 (FFAR1) using virtual screening. *Journal of Medicinal Chemistry*, 51(3):625–633.

[Tsai et al., 1999] Tsai, J., Taylor, R., Chothia, C., and Gerstein, M. (1999). The packing density in proteins: standard radii and volumes. *Journal of Molecular Biology*, 290(1):253–266.

[Tyagi et al., 2008] Tyagi, M., de Brevern, A. G., Srinivasan, N., and Offmann, B. (2008). Protein structure mining using a structural alphabet. *Proteins*, 71(2):920–937.

[Van De Water et al., 2004] Van De Water, T. R., Lallemend, F., Eshraghi, A. A., Ahsan, S., He, J., Guzman, J., Polak, M., Malgrange, B., Lefebvre, P. P., Staecker, H., and Balkany, T. J. (2004). Caspases, the enemy within, and their role in oxidative stress-induced apoptosis of inner ear sensory cells. *Otology & Neurotology*, 25(4):627–632.

[Vastrik et al., 2007] Vastrik, I., D'Eustachio, P., Schmidt, E., Joshi-Tope, G., Gopinath, G., Croft, D., de Bono, B., Gillespie, M., Jassal, B., Lewis, S., Matthews, L., Wu, G., Birney, E., and Stein, L. (2007). Reactome: a knowledgebase of biological pathways and processes. *Genome Biology*, 8:R39+.

[Venter et al., 2001] Venter, J. C., Adams, M. D., Myers, E. W., Li, P. W., Mural, R. J., Sutton, G. G., Smith, H. O., Yandell, M., Evans, C. A., Holt, R. A., Gocayne, J. D., Amanatides, P., Ballew, R. M., Huson, D. H., Wortman, J. R., Zhang, Q., Kodira, C. D., Zheng, X. H., Chen, L., Skupski, M., Subramanian, G., Thomas, P. D., Zhang, J., Gabor Miklos, G. L., Nelson, C., Broder, S., Clark, A. G., Nadeau, J., McKusick, V. A., Zinder, N., Levine, A. J., Roberts, R. J., Simon, M., Slayman, C., Hunkapiller, M., Bolanos, R., Delcher, A., Dew, I., Fasulo, D., Flanigan, M., Florea, L., Halpern, A., Hannenhalli, S., Kravitz, S., Levy, S., Mobarry, C., Reinert, K., Remington, K., Abu-Threideh, J., Beasley, E., Biddick, K., Bonazzi, V., Brandon, R., Cargill, M., Chandramouliswaran, I., Charlab, R., Chaturvedi, K., Deng, Z., Di Francesco, V., Dunn, P., Eilbeck, K., Evangelista, C., Gabrielian, A. E., Gan, W., Ge, W., Gong, F., Gu, Z., Guan, P., Heiman, T. J., Higgins, M. E., Ji, R. R., Ke, Z., Ketchum, K. A., Lai, Z., Lei, Y., Li, Z., Li, J., Liang, Y., Lin, X., Lu, F., Merkulov, G. V., Milshina, N., Moore, H. M., Naik, A. K., Narayan, V. A., Neelam, B., Nusskern, D., Rusch, D. B., Salzberg, S., Shao, W., Shue, B., Sun, J., Wang, Z., Wang, A., Wang, X., Wang, J., Wei, M., Wides, R., Xiao, C., Yan, C., Yao, A., Ye, J., Zhan, M., Zhang, W., Zhang, H., Zhao, Q., Zheng, L., Zhong, F., Zhong, W., Zhu, S., Zhao, S., Gilbert, D., Baumhueter, S., Spier, G., Carter, C., Cravchik, A., Woodage, T., Ali, F., An, H., Awe, A., Baldwin, D., Baden, H., Barnstead, M., Barrow, I., Beeson, K., Busam, D., Carver, A., Center, A., Cheng, M. L., Curry, L., Danaher, S., Davenport, L., Desilets, R., Dietz, S., Dodson, K., Doup, L., Ferriera, S., Garg, N., Gluecksmann, A., Hart, B., Haynes, J., Haynes, C., Heiner, C., Hladun, S., Hostin, D., Houck, J., Howland, T., Ibegwam, C., Johnson, J., Kalush, F., Kline, L., Koduru, S., Love, A., Mann, F., May, D., McCawley, S., McIntosh, T., McMullen, I., Moy, M., Moy, L., Murphy, B., Nelson, K., Pfannkoch, C., Pratts, E., Puri, V., Qureshi, H., Reardon, M., Rodriguez, R., Rogers, Y. H., Romblad, D., Ruhfel, B., Scott, R., Sitter, C., Smallwood, M., Stewart, E., Strong, R., Suh, E., Thomas, R., Tint, N. N., Tse, S., Vech, C., Wang, G., Wetter, J., Williams, S., Williams, M., Windsor, S., Winn-Deen, E., Wolfe, K., Zaveri, J., Zaveri, K., Abril, J. F., Guigó, R., Campbell, M. J., Sjolander, K. V., Karlak, B., Kejariwal, A., Mi, H., Lazareva, B., Hatton, T., Narechania, A., Diemer, K., Muruganujan, A., Guo, N., Sato, S., Bafna, V., Istrail, S., Lippert, R., Schwartz, R., Walenz, B., Yooseph, S., Allen, D., Basu, A., Baxendale, J., Blick, L., Caminha, M., Carnes-Stine, J., Caulk, P., Chiang, Y. H., Coyne, M., Dahlke, C., Mays, A., Dombroski, M., Donnelly, M., Ely, D., Esparham, S., Fosler, C., Gire, H., Glanowski, S., Glasser, K., Glodek, A., Gorokhov, M., Graham, K., Gropman, B., Harris, M., Heil, J., Henderson, S., Hoover, J., Jennings, D., Jordan, C., Jordan, J., Kasha, J., Kagan, L., Kraft, C., Levitsky, A., Lewis, M., Liu, X., Lopez, J., Ma, D., Majoros, W., McDaniel, J., Murphy, S., Newman, M., Nguyen, T., Nguyen, N., Nodell, M., Pan, S., Peck, J., Peterson, M., Rowe, W., Sanders, R., Scott, J., Simpson, M., Smith, T., Sprague, A., Stockwell, T., Turner, R., Venter, E., Wang, M., Wen, M., Wu, D., Wu, M., Xia, A., Zandieh, A., and Zhu, X. (2001). The sequence of the human genome. *Science*, 291(5507):1304–1351.

[Voigt et al., 2001] Voigt, J. H., Bienfait, B., Wang, S., and Nicklaus, M. C. (2001). Comparison of the NCI open database with seven large chemical structural databases. *Journal of Chemical Information and Computer Sciences*, 41(3):702–712.

[Wadley et al., 2007] Wadley, L. M., Keating, K. S., Duarte, C. M., and Pyle, A. M. (2007). Evaluating and Learning from RNA Pseudotorsional Space: Quantitative Validation of a Reduced Representation for RNA Structure. *Journal of Molecular Biology*, 372(4):942–957.

[Wang and Dunbrack, 2005] Wang, G. and Dunbrack, R. L. (2005). PISCES: recent improvements to a PDB sequence culling server. *Nucleic Acids Resarch*, 33(Web Server issue):W94–W98.

[Wang et al., 2005] Wang, R., Fang, X., Lu, Y., Yang, C. Y., and Wang, S. (2005). The PDBbind database: methodologies and updates. *Journal of Medicinal Chemistry*, 48(12):4111–4119.

[Watson and Crick, 1953] Watson, J. D. and Crick, F. H. (1953). Molecular structure of nucleic acids: A structure for deoxyribose nucleic acid. *Nature*, 171(4356):737–738.

[Weinger et al., 2004] Weinger, J. S., Parnell, M. M., Dorner, S., Green, R., and Strobel, S. A. (2004). Substrate-assisted catalysis of peptide bond formation by the ribosome. *Nature Structural & Molecular Biology*, 11(11):1101+.

[Wheeler et al., 2008] Wheeler, D. L., Barrett, T., Benson, D. A., Bryant, S. H., Canese, K., Chetvernin, V., Church, D. M., Dicuccio, M., Edgar, R., Federhen, S., Feolo, M., Geer, L. Y., Helmberg, W., Kapustin, Y., Khovayko, O., Landsman, D., Lipman, D. J., Madden, T. L., Maglott, D. R., Miller, V., Ostell, J., Pruitt, K. D., Schuler, G. D., Shumway, M., Sequeira, E., Sherry, S. T., Sirotkin, K., Souvorov, A., Starchenko, G., Tatusov, R. L., Tatusova, T. A., Wagner, L., and Yaschenko, E. (2008). Database resources of the National Center for Biotechnology Information. *Nucleic Acids Research*, 36(Database issue):D13–D21.

[Whittle et al., 2003] Whittle, M., Willett, P., Klaffke, W., and van Noort, P. (2003). Evaluation of similarity measures for searching the dictionary of natural products database. *Journal of Chemical Information and Computer Sciences*, 43(2):449–457.

[Winkler et al., 2002] Winkler, W. C., Cohen-Chalamish, S., and Breaker, R. R. (2002). An mRNA structure that controls gene expression by binding FMN. *Proceedings of the National Academy of Sciences*, 99(25):15908–15913.

[Wishart et al., 2008] Wishart, D. S., Knox, C., Guo, A. C., Cheng, D., Shrivastava, S., Tzur, D., Gautam, B., and Hassanali, M. (2008). DrugBank: a knowledgebase for drugs, drug actions and drug targets. *Nucleic Acids Research*, 36(Database issue):D901–D906.

[Xiao and Rajewsky, 2009] Xiao, C. and Rajewsky, K. (2009). MicroRNA control in the immune system: basic principles. *Cell*, 136(1):26–36.

[Yang et al., 2003] Yang, H., Jossinet, F., Leontis, N., Chen, L., Westbrook, J., Berman, H., and Westhof, E. (2003). Tools for the automatic identification and classification of RNA base pairs. *Nucleic Acids Research*, 31(13):3450–3460.

[Yengi, 2005] Yengi, L. G. (2005). Systems biology in drug safety and metabolism: integration of microarray, real-time PCR and enzyme approaches. *Pharmacogenomics*, 6(2):185–192.

[Yeturu and Chandra, 2008] Yeturu, K. and Chandra, N. (2008). PocketMatch: A new algorithm to compare binding sites in protein structures. *BMC Bioinformatics*, 9(1):543+.

[Yeung et al., 2008] Yeung, N., Cline, M. S., Kuchinsky, A., Smoot, M. E., and Bader, G. D. (2008). Exploring biological networks with Cytoscape software. *Current Protocols in Bioinformatics*, Chapter 8(Unit 8.13).

[Yu et al., 2008] Yu, H., Braun, P., Yildirim, M. A., Lemmens, I., Venkatesan, K., Sahalie, J., Hirozane-Kishikawa, T., Gebreab, F., Li, N., Simonis, N., Hao, T., Rual, J.-F., Dricot, A., Vazquez, A., Murray, R. R., Simon, C., Tardivo, L., Tam, S., Svrzikapa, N., Fan, C., de Smet, A.-S., Motyl, A., Hudson, M. E., Park, J., Xin, X., Cusick, M. E., Moore, T., Boone, C., Snyder, M., Roth, F. P., Barabasi, A.-L., Tavernier, J., Hill, D. E., and Vidal, M. (2008). High-Quality Binary Protein Interaction Map of the Yeast Interactome Network. *Science*, 322(5898):104–110.

[Zhang and Skolnick, 2005] Zhang, Y. and Skolnick, J. (2005). TM-align: a protein structure alignment algorithm based on the TM-score. *Nucleic Acids Research*, 33(7):2302–2309.

[Zhang and Skolnick, 2007] Zhang, Y. and Skolnick, J. (2007). Scoring function for automated assessment of protein structure template quality. *Proteins*, 57(4):702–710.

Document specifications

Version:	Final
Compilation Date:	December 13, 2009