Dissertation zur Erlangung des Doktorgrades eingereicht am
Fachbereich Mathematik und Informatik der Freien Universität Berlin

The Common Denominator Procedure

A Novel Approach to Gene Expression Data Mining for Identification of Phenotype-Specific Genes

René Korn

2006

Betreuer:
PD Dr. Steffen Schulze-Kremer
Dr. Sascha Röhrig
Dr. Ulrich Brinkmann

Gutachter:
PD Dr. Steffen Schulze-Kremer
Prof. Dr. Peter Buckel

Tag der Disputation:
28. April 2006

Dedicated to my beloved father Klaus Korn,
who had to suffer from cancer for too many years.

Abstract

This thesis addresses the gap between the amount of on-hand expression data and the availability of information related to the function of those genes. To this end, a data mining procedure for the identification of genes that are associated with pre-defined phenotypes and/or molecular pathways was established. Based on the observation that pathway/phenotype associated genes are frequently expressed in same or nearby places and at identical or similar time points, an approach termed COMMON DENOMINATOR PROCEDURE (CDP) was devised. One unique feature of this novel approach is that the specificity and probability to identify desired phenotype/pathway-associated factors increases the more diverse the input data are. Three different approaches are discussed and compared: (i) a BASIC CDP, (ii) a GENETIC ALGORITHM BASED CDP and (iii) an INDICATOR GENES BASED CDP. To show the feasibility of these approaches, the CGAP Expression Data combined with a defined set of angiogenic factors was used to identify additional and novel angiogenesis-associated genes. A multitude of these additional genes were known to be associated with angiogenesis according to published data, verifying our approach. Application of a high throughput functional genomics platform (XantoScreen™) provided further experimental evidence for association of candidate genes with angiogenesis.

Acknowledgement

First of all, I would like to thank PD Dr. Steffen Schulze-Kremer for being my supervisor and for supporting this work with his beneficial ideas and constructive criticism. I would also like to kindly thank Prof. Dr. Peter Buckel for taking over the part of the second referee. I would like to thank Dr. Ulrich Brinkmann for the original impulse to this work and his commitment.

I also appreciate all my friends and colleagues from Xantos, for their helpful feedback and their friendship. I want to thank Dr. Irene Boche, Stefan Hess, Piia Lämmlein, Dr. Dieter Link, Dr. Stephan Reschauer and Dr. Jürgen Zitzler for their friendship, a nice time and exciting badminton matches. Special thanks go to Dr. Sascha Röhrig, my patient superior, mentor and friend, as well as my closest colleagues from the Bioinformatics and IT department, Dr. Bettina Ehring, Dr. Beate Gawin, Fabian Weiß, Alexander Felber, Dr. Björn Kesper, Dr. Reinhold Köckerbauer and Dr. Alexander Spychaj. I am particularly grateful for the support from the high throughput screening team, Johannes Görl, Dr. Kerstin König-Hoffmann, Dr. Rolf Schäfer and Michael Kazinski.

I want to demonstrate my deepest respect to Xantos Biomedicine AG, especially to Stephan Wehselau and again to Prof. Dr. Peter Buckel, for there investment in my future and their commitment during exciting, interesting and turbulent times.

I am also thankful for the possibility to submit my work to the Department of Mathematics and Computer Science at the Free University of Berlin.

Last but not least I want to thank my family, especially my partner in life Janine Christ, my mother Brigitte Korn, my sister Carolin Korn and my godfather Rainer Spitzenpfeil as well as all my friends, especially Roman Egle and Daniel Godau, for encouragement, for proofreading, for putting up with me all the time, and for just being there.

Contents

List of Figures

List of Tables

Abbreviations

bCDP	Basic Common Denominator Procedure
cDNA	Complementary DNA
CDP	Common Denominator Procedure
CGAP	NCI's Cancer Genome Anatomy Project
CPU	Central Processing Unit
DDIT4	HIF-1 Responsive RTP801
DMEM	Dulbecco's Modified Eagle Medium
DNA	Deoxyribonucleic Acid
ECBM	Endothelial Cell Basal Medium, Promocell®
ECGF	Endothelial Cell Growth Factor 1
ECGM	Endothelial Cell Growth Medium, Promocell®
EPAS1	Endothelial PAS Domain Protein 1
ERD	Entity Relationship Diagram
EST	Expressed Sequence Tag
FDA	U. S. Food and Drug Administration
FCS	Foetal Calf Serum, Invitrogen®
GA	Genetic Algorithm
GB	Gigabyte
GHz	Gigahertz
gaCDP	Genetic Algorithm Based Common Denominator Procedure
GO	Gene Ontology
GRIF	Gene References Into Function
HEK293	Human Embryonal Kidney Cells
HIF1A	Hypoxia-inducible Factor 1, Alpha Subunit
HUVEC	Human Umbilical Cord Vein Endothelial Cells
igCDP	Indicator Genes Based Common Denominator Procedure
IGFR1	Insulin-like Growth Factor 1 Receptor
JGAP	Java Genetics Algorithms Package
MB	Megabyte
mRNA	Messenger RNA
NCBI	U. S. National Center for Biotechnology Information
NCI	U. S. National Cancer Institute
NIH	U. S. National Health Institute
NHGRI	U. S. National Human Genome Research Institute
PDGF	Platelet-derived Growth Factor

RAM	Random Access Memory
RNA	Ribonucleic Acid
SAGE	Serial Analysis of Gene Expression
VEGF	Vascular Endothelial Growth Factor
UniProt	Universal Protein Resource
RefSeq	Protein Sequence Databases (Reference Sequence)
GenBank	Nucleotide Sequence Database
EMBL	European Molecular Biology Laboratory
UML	Unified Modeling Language