DISSERTATIONES MATHEMATICAE UNIVERSITATIS TARTUENSIS
74

DISSERTATIONES MATHEMATICAE UNIVERSITATIS TARTUENSIS
74

MEELIS KULL

Statistical enrichment analysis in algorithms for studying gene regulation

TARTU UNIVERSITY PRESS

Institute of Computer Science, Faculty of Mathematics and Computer Science, University of Tartu, Estonia

Dissertation accepted for public defense of the degree of Doctor of Philosophy (PhD) on June 29, 2011 by the Council of the Institute of Computer Science, University of Tartu.

Supervisor:

Prof. Jaak Vilo
University of Tartu
Tartu, Estonia

Opponents:

Dr. Joaquín Dopazo
Príncipe Felipe Research Centre
Valencia, Spain

Dr. Juho Rousu
University of Helsinki
Helsinki, Finland

The public defense will take place on August 26, 2011 at 15:00 in Liivi 2-403.

The publication of this dissertation was financed by Institute of Computer Science, University of Tartu.

ISSN 1024-4212
ISBN 978-9949-19-804-7 (trükis)
ISBN 978-9949-19-805-4 (PDF)

Tartu Ülikooli Kirjastus
http://www.tyk.ee
Tellimus nr. 462

Contents

List of Original Publications **7**

Abstract **8**

Introduction **9**

1 Preliminaries and notations **12**
1.1 Biological preliminaries 12
1.2 Statistical preliminaries 14
1.3 Notations 20

2 Statistical enrichment analysis **21**
2.1 Definition of enrichment 21
2.2 Functional enrichment analysis 24
2.3 Regulatory enrichment analysis 24
2.4 Association between properties 26

3 Hierarchical clustering and functional enrichment **29**
3.1 Motivation 29
3.2 Fast approximate hierarchical clustering (Paper I) 30
3.3 Hierarchical functional enrichment analysis of microarray data (Paper II) 31

4 Regulatory enrichment in promoter analysis **33**
4.1 Motivation 33
4.2 The proposed method for measuring association 34
4.3 Computational experiments 38
4.4 Promoter analysis of genes differentially expressed in mouse mesoderm development (in Paper III) 40
4.5 Promoter analysis of genes differentially expressed in mouse adipogenesis (in Paper IV) 42

5 Evolutionary models of enrichment (Paper V) 44
5.1 Motivation 44
5.2 The proposed evolutionary model 45
5.3 Discussion 46

Conclusions 48

A Proofs of Theorems 4.1 and 4.2 50

Bibliography 57

Acknowledgments 64

Kokkuvõte (Summary in Estonian) 65

Original Publications 67
Fast approximate hierarchical clustering using similarity heuristics . . . 71
VisHiC–hierarchical functional enrichment analysis of microarray data . 87
Global transcriptomic analysis of murine embryonic stem cell-derived brachyury^{+} (T) cells 95
Comprehensive transcriptome analysis of mouse embryonic stem cell adipogenesis unravels new processes of adipocyte development . . 117
An evolutionary model of DNA substring distribution 135

Curriculum Vitae 146

Elulookirjeldus 147

DISSERTATIONES MATHEMATICAE UNIVERSITATIS TARTUENSIS 148

LIST OF ORIGINAL PUBLICATIONS

1. Kull, M., Vilo, J.: Fast approximate hierarchical clustering using similarity heuristics. BioData Mining 1(1), 9 (Sep 2008).

2. Krushevskaya, D., Peterson, H., Reimand, J., Kull, M., Vilo, J.: VisHiC – hierarchical functional enrichment analysis of microarray data. Nucl. Acids Res. 37(Web Server issue), W587–92 (Jul 2009).

3. Doss, M.X., Wagh, V., Schulz, H., Kull, M., Kolde, R., Pfannkuche, K., Nolden, T., Himmelbauer, H., Vilo, J., Hescheler, J., Sachinidis, A.: Global transcriptomic analysis of murine embryonic stem cell-derived brachyury^{+} (T) cells. Genes to Cells 15(3), 209–228 (Feb 2010).

4. Billon, N., Kolde, R., Reimand, J., Monteiro, M.C., Kull, M., Peterson, H., Tretyakov, K., Adler, P., Wdziekonski, B., Vilo, J., Dani, C.: Comprehensive transcriptome analysis of mouse embryonic stem cell adipogenesis unravels new processes of adipocyte development. Genome Biol 11(8), R80 (Aug 2010).

5. Kull, M., Tretyakov, K., Vilo, J.: An evolutionary model of DNA substring distribution. In: Elomaa, T., Mannila, H., Orponen, P. (eds.) Algorithms and Applications, Essays Dedicated to Esko Ukkonen on the Occasion of His 60th Birthday, Lecture Notes in Computer Science, vol. 6060, pp. 147–157. Springer (2010).

ABSTRACT

Statistical enrichment analysis is a family of data analysis methods studying whether the data are enriched in some quantity, and by how much. The analysis can be applied when we have some expectation about the numeric value of the quantity. Then enrichment refers to a situation where the actual value turns out to be significantly higher than expected. Enrichment analysis has been used extensively in bioinformatics for studying associations between biological entities (such as genes, biological processes, cellular components, molecular functions, signalling pathways, regulatory mechanisms) by combining the data from experiments and biological databases.

The goal of this dissertation is to enhance and apply algorithms involving or related to statistical enrichment analysis for studying gene regulation. The major contributions of the dissertation are the following.

- First, a formal statistical definition of enrichment is proposed, complemented by the presentation of several known enrichment analysis methods with respect to the new definition.
- Second, a fast approximate algorithm is developed for performing hierarchical clustering. This is applied in a software tool for performing hierarchical functional enrichment analysis of gene expression data, suitable as one of the first steps in studying gene regulation.
- Third, a novel measure of enrichment strength is developed in the context of regulatory enrichment analysis, which is a proposed extension of motif enrichment analysis. The new measure is applied in two biological studies of gene regulation in mouse embryonic stem cells.
- Finally, an evolutionary DNA substring distribution model is proposed with potential applications in background modelling for motif discovery and motif enrichment analysis.

INTRODUCTION

As more and more genomes of different organisms are sequenced, we are starting to have a pretty good overview of the variety of genes, the main building instructions of life on Earth. While the genes encoded in DNA specify *how to build* RNA and proteins, they do not provide the information about *when and how much to build*. *Gene regulation* is a term referring to a long list of mechanisms that affect the timing and production rate of gene products. Disruption in regulation of a single gene is a cause for multiple syndromes and diseases in human [28].

The major sources of information about gene regulation are biological high-throughput experiments. The task of bioinformaticians is to analyze such large data sets, draw conclusions and propose hypotheses, which can later be verified or disproved in further experiments. In addition to developing new methods, bioinformatics applies and combines methods originating from statistics, algorithmics, machine learning, data mining.

The biological focus of this dissertation is on transcriptional gene regulation. We propose several algorithms for analyzing data about gene expression, regulatory sequences and functional annotations of genes. More specifically, the proposed methods are related to studying the associations between these types of data by means of statistical enrichment analysis.

Enrichment, or higher value of a quantity than expected according to some reference, is a notion that has been used broadly in science. The analyses searching for regulatory signals or *motifs* in the genome were among the first applications in bioinformatics [13, 49]. The emerging technologies for high-throughput gene expression measurements provided new data which was incorporated in the enrichment-based motif discovery methods [11, 29]. Since then, statistical enrichment or *over-representation* of motifs has become a wide-spread method for studying transcriptional gene regulation [17, 45]. The increasing knowledge and data about the transcription regulatory mechanisms [23, 25] help to focus the search for regulatory motifs on appropriate genomic regions and open up further perspectives in integrating multiple types of data about gene regulation [63].

Another important application of enrichment in bioinformatics is *functional enrichment analysis*. With systematic collection of knowledge about the function

of genes into databases [3, 47] it became possible to study if an experimentally derived gene set is enriched in genes with the same functional annotation [21, 61]. Functional enrichment analysis has become a standard technique in interpreting gene expression data [37] and gene sets obtained from any experimental and computational protocols [32].

This dissertation proposes a general theoretical framework for statistical enrichment analysis. The main contributions of the dissertation are the proposed algorithms and analysis methods for studying gene regulation, all related to enrichment to some extent. These algorithms and methods have been published in the following five papers, with the contribution of the author of the dissertation highlighted.

Paper I Kull, M., Vilo, J.: Fast approximate hierarchical clustering using similarity heuristics. BioData Mining 1(1), 9 (Sep 2008).

The problem statement and background information for this publication were provided by the supervisor J. Vilo, everything else is by M. Kull, the author of this dissertation.

Paper II Krushevskaya, D., Peterson, H., Reimand, J., Kull, M., Vilo, J.: VisHiC – hierarchical functional enrichment analysis of microarray data. Nucl. Acids Res. 37(Web Server issue), W587–92 (Jul 2009).

The problem statement is by J. Vilo and the final web server was developed by D. Krushevskaya. M. Kull built the first prototype of the software tool and took part in the discussions.

Paper III Doss, M.X., Wagh, V., Schulz, H., Kull, M., Kolde, R., Pfannkuche, K., Nolden, T., Himmelbauer, H., Vilo, J., Hescheler, J., Sachinidis, A.: Global transcriptomic analysis of murine embryonic stem cell-derived brachyury$^+$ (T) cells. Genes to Cells 15(3), 209–228 (Feb 2010).

M. Kull developed a method for and carried out the promoter analysis, and wrote the description of the analysis for the paper.

Paper IV Billon, N., Kolde, R., Reimand, J., Monteiro, M.C., Kull, M., Peterson, H., Tretyakov, K., Adler, P., Wdziekonski, B., Vilo, J., Dani, C.: Comprehensive transcriptome analysis of mouse embryonic stem cell adipogenesis unravels new processes of adipocyte development. Genome Biol 11(8), R80 (Aug 2010).

H. Peterson complemented the promoter analysis method from Paper III with the criteria for evolutionary conservation and the analysis was performed jointly by H. Peterson, M. Kull and K. Tretyakov. M. Kull wrote the description of the analysis for the paper.

Paper V Kull, M., Tretyakov, K., Vilo, J.: An evolutionary model of DNA substring distribution. In: Elomaa, T., Mannila, H., Orponen, P. (eds.) Algorithms and Applications, Essays Dedicated to Esko Ukkonen on the Occasion of His 60th Birthday, Lecture Notes in Computer Science, vol. 6060, pp. 147–157. Springer (2010).

M. Kull proposed the problem statement and wrote the first draft of the paper. The experiments were performed and final paper was written jointly by M. Kull and K. Tretyakov.

The copies of papers I–V are included at the end of the dissertation on pp.67–145.

The outline of the dissertation is the following. Chapter 1 provides the biological and statistical preliminaries and introduces the notations. Chapter 2 gives a formal definition of statistical enrichment analysis and describes two of its well-known applications in bioinformatics – functional enrichment analysis and regulatory enrichment analysis. The rest of the chapters introduce the work done for Papers I–V. Chapter 3 proposes an algorithm for fast approximate hierarchical clustering which is used in a web server developed for visualizing gene expression data together with the results of hierarchical functional enrichment analysis. Chapter 4 proposes a novel measure of enrichment strength in the context of gene promoter analysis and describes the results of its application in two biological studies. Chapter 5 proposes an evolutionary DNA substring distribution model with potential applications in background modelling for motif discovery and motif enrichment analysis. Finally, Appendix A contains the proofs of Theorems 4.1 and 4.2 which are stated in Chapter 4.

CHAPTER 1

PRELIMINARIES AND NOTATIONS

In this chapter we provide the biological preliminaries of the dissertation as well as the statistical preliminaries and notations used in Chapters 2 and 4 and in Appendix A.

1.1 Biological preliminaries

Genetic and epigenetic information

All living organisms on earth contain genetic information encoded in long molecules called *nucleic acids – ribonucleic acid (RNA)* and *deoxyribonucleic acid (DNA)*. DNA stands as a long-term memory with instructions for producing RNA and proteins, which in turn catalyze most of the chemical reactions and serve several other functions in the living cells. The nucleic acids consist of a sugar-phosphate backbone with a sequence of nucleotides attached to it. Three nucleotides – adenine (A), cytosine (C), guanine (G) – are in common for DNA and RNA, whereas the fourth is different: thymine (T) for DNA and uracil (U) for RNA. The sequence of nucleotides is the best known mechanism for information coding in the living organisms and is called the *genetic information*. However, several other mechanisms for storing long-term information have been discovered recently, called together as *epigenetic information* [10]. Epigenetic information provides landmarks to guide the molecules that interact with DNA and modifies how the genetic information is interpreted [23]. The major forms of epigenetic information include the methylation of nucleotides in the DNA, the location of *nucleosomes*, which are the packaging units of DNA, and chemical modification in *histones*, which are the proteins involved in the formation of nucleosomes.

Transcription

Transcription is the synthesis of RNA molecules using DNA as a template. It is catalyzed by the protein *RNA polymerase* which moves along the DNA and attaches the nucleotide which is *complementary* to the current DNA nucleotide to the end of the RNA molecule. The complementary nucleotides for A, C, G, T in DNA are U, G, C, A in RNA, respectively.

Genetic and epigenetic information becomes useful for the cell through transcription. The RNA molecules resulting from transcription become mature by going through post-transcriptional modification and after this have a multitude of functions. The most widely known function is performed by the messenger RNA (mRNA) which is transported out of the nucleus and is used as a template for *translation* – the synthesis of poly-peptides which are folded into proteins. Translation is supported by another type of RNA molecules called transfer RNAs (tRNA). MicroRNAs (miRNA) selectively lead other RNA molecules to degradation.

Genes and gene regulation

Gene is a genomic sequence directly encoding functional product molecules [58]. The process by which the information encoded in the gene is used to produce the gene product is called *gene expression*. *Gene regulation* or *regulation of gene expression* is the process which determines the timing and quantity of gene expression.

In principle, gene regulation can act at any of the multiple steps of gene expression. For example, the regulated steps include transcription, post-transcriptional modification and translation. Probably the best known regulators are *transcription factors* – proteins which bind DNA and by this affect the rate of transcription.

Transcription regulation

The crucial genomic feature in the regulation of transcription is the *promoter*. This is the region where the RNA polymerase attaches and starts transcription. Transcription factors can affect the rate of RNA polymerase binding to the promoter and transcribing the gene. Most transcription factors bind in a sequence-specific manner to *transcription factor binding sites* in the genome, whereas the binding is guided by epigenetic information [23]. These binding sites can be located in the promoter but also further away in the regions called *enhancers*, which get into contact with the promoter region by DNA looping. A transcription factor is said to regulate a gene, if its binding modulates the transcription rate in some biological condition.

1.2 Statistical preliminaries

Here we give a simplified view of the main concepts in probability theory, enough to understand the dissertation. For an axiomatic probability theory refer to Billingsley [8].

Random variable, probability and distribution

A *random variable* is any function with the *sample space* as its domain. The sample space includes all possible outcomes of the stochastic experiments we reason about, whereas the random variables highlight the features we are interested in. In the dissertation we use only real-valued random variables, and sometimes restrict to binary or discrete values ($\{0, 1\}$, $\mathbb{N}$). In particular, X and Y denote binary random variables throughout the dissertation.

Probability measure $\mathbf{P}$ is a function assigning probabilities (values in the range $[0, 1]$) to subsets of the sample space which are called *events*. Events are commonly represented as predicates including random variables. For example, $\mathbf{P}(X{=}1,\ Y{=}0)$ denotes the probability of the event of $X{=}1$ and $Y{=}0$ occurring simultaneously. We say that an event occurs *almost surely* (a.s.), if its probability is 1.

A function $\mathcal{D}$ is the *distribution* of a random variable W, if it specifies the probabilities of W taking on values from various sets $\mathcal{S}$, that is $\mathcal{D}(\mathcal{S}) = \mathbf{P}(W{\in}\mathcal{S})$. In such case we say that W is distributed as $\mathcal{D}$, and denote it by $W \sim \mathcal{D}$. Several random variables can have the same distribution and a distribution can even be defined without specifying any random variable. For any real-valued random variable W its distribution is uniquely determined by the *cumulative distribution function* F^{W} defined as

$$F^{\mathsf{W}}(w) = \mathbf{P}(W{\leq}w)$$

The distribution of a binary random variable X is uniquely determined by the probability $\mathbf{P}(X{=}1)$. In order to achieve the coherence of notations with the empirical probabilities defined below, we denote

$$P^{\mathsf{X}}(x) = \mathbf{P}(X{=}x)$$

The *joint distribution* of a set of random variables specifies the probabilities of all combinations of outputs of these variables. In particular, the joint distribution of binary random variables X and Y specifies for any $x, y \in \{0, 1\}$ the probability

$$P^{\mathsf{XY}}(x,\ y) = \mathbf{P}(X{=}x,\ Y{=}y)$$

Probability distributions used in the dissertation

A binary random variable X has *Bernoulli distribution* with parameter p, denoted as $X \sim \mathcal{B}(1, p)$ if

$$\mathbf{P}(X{=}x) = \begin{cases} p & \text{if } x = 1, \\ 1 - p & \text{if } x = 0. \end{cases}$$

Every binary random variable has Bernoulli distribution with some parameter p.

A random variable Z has *hypergeometric distribution* which we denote as $Z \sim \mathcal{H}(n, k, m)$, if $k, m, n \in \mathbb{N}$ and $0 \leq k, m \leq n$ and for any $z \in \mathbb{N}$ the following holds:

$$\mathbf{P}(Z{=}z) = \frac{\binom{m}{z}\binom{n-m}{k-z}}{\binom{n}{k}}$$

The hypergeometric distribution models a situation where we are counting the number of marked balls obtained while randomly drawing k balls without replacement from a box with m marked balls and n balls in total.

A random variable W has *normal distribution*, denoted as $W \sim \mathcal{N}(\mu, \sigma^2)$ if

$$\mathbf{P}(W{\leq}w) = \int_{-\infty}^{w} \frac{1}{\sqrt{2\pi\sigma^2}} e^{-\frac{(w-\mu)^2}{2\sigma^2}} dw$$

Normal distribution is important in many contexts in statistics, for example the sum of many independent and identically distributed random variables approaches normal distribution.

Conditional probability and independence

Conditional probability $\mathbf{P}(A|B)$ is the probability of the event A given that the event B has occurred, and can be calculated as

$$\mathbf{P}(A|B) = \frac{\mathbf{P}(A,\, B)}{\mathbf{P}(B)}$$

Conditional probability is defined only if $\mathbf{P}(B) > 0$. A function $\mathcal{D}$ is the *conditional distribution* of a random variable W given an event A, if it specifies the probabilities of W taking on values from various sets $\mathcal{S}$ given that the event A has occurred, that is $\mathcal{D}(\mathcal{S}) = \mathbf{P}(W \in \mathcal{S}|A)$. In such case we say that $W|A$ is distributed as $\mathcal{D}$ and denote it by $W|A \sim \mathcal{D}$.

For binary random variables X and Y the conditional distribution of $Y|X{=}x$ specifies the probabilities

$$P^{\mathsf{Y}|\mathsf{X}}(y|x) = \mathbf{P}(Y{=}y|X{=}x)$$

for $y = 0, 1$. For a binary random variable X and a real-valued random variable W the conditional distribution of $W|X{=}x$ specifies the *conditional cumulative distribution function* of W given $X{=}x$ as a function of $w \in \mathbb{R}$:

$$F^{\mathsf{W}|\mathsf{X}}(w|x) = \mathbf{P}(W{\leq}w|X{=}x)$$

Conditional probabilities and distributions provide intuition to understand the notion of *independence*. Events A and B are called *independent*, if

$$\mathbf{P}(A,\, B) = \mathbf{P}(A)\,\mathbf{P}(B)$$

The independence of A and B is equivalent to $\mathbf{P}(A|B) = \mathbf{P}(A)$ assuming $\mathbf{P}(B) > 0$. Random variables W and W' are called independent, if the events $\{W \in \mathcal{S}\}$ and $\{W' \in \mathcal{S}'\}$ are independent for any sets $\mathcal{S}$ and $\mathcal{S}'$. For a real-valued W it is enough to consider the events $\{W \leq w\}$ for all $w \in \mathbb{R}$ instead of all events $\{W \in \mathcal{S}\}$. For a binary random variable X it is enough to consider the event $\{X{=}1\}$ instead of all events $\{X \in \mathcal{S}\}$.

In particular, the binary random variables X and Y are independent if the events $\{X{=}1\}$ and $\{Y{=}1\}$ are independent, that is

$$P^{\mathsf{XY}}(1,\, 1) = P^{\mathsf{X}}(1)\, P^{\mathsf{Y}}(1)$$

If $0 < P^{\mathsf{X}}(1) < 1$, then this is equivalent to

$$P^{\mathsf{Y}|\mathsf{X}}(1|1) = P^{\mathsf{Y}|\mathsf{X}}(1|0) = P^{\mathsf{Y}}(1)$$

A binary random variable X and a real-valued random variable W are independent, if the events $\{X{=}1\}$ and $\{W{\leq}w\}$ are independent for any $w \in \mathbb{R}$. If $0 < P^{\mathsf{X}}(1) < 1$, then this is equivalent to

$$F^{\mathsf{W}|\mathsf{X}}(w|1) = F^{\mathsf{W}|\mathsf{X}}(w|0) = F^{\mathsf{W}}(w) \quad \forall w \in \mathbb{R}$$

Non-independent random variables are said to be *dependent* or *associated*.

Events A and B are called *conditionally independent* given an event C if

$$\mathbf{P}(A,\, B|C) = \mathbf{P}(A|C)\,\mathbf{P}(B|C)$$

A binary random variable X and a real-valued random variable W are conditionally independent given a binary random variable Y, if for any $x, y \in \{0, 1\}$ and $w \in \mathbb{R}$ the events $\{X{=}x\}$ and $\{W{\leq}w\}$ are conditionally independent given the event $\{Y{=}y\}$.

Measures of association

In this dissertation we use three different measures for quantifying the association between dependent binary random variables. There exist many other similar measures, for a review refer to Tan *et al.* [60] or Huynh *et al.* [33].

Pearson correlation, also known as the Pearson product-moment correlation coefficient, between binary random variables X and Y is defined as follows:

$$\operatorname{corr}(X,Y) = \frac{\mathbf{E}(XY) - \mathbf{E}(X)\cdot\mathbf{E}(Y)}{\sqrt{\mathbf{D}(X)\cdot\mathbf{D}(Y)}} = \frac{P^{\mathsf{XY}}(1,\,1) - P^{\mathsf{X}}(1)\,P^{\mathsf{Y}}(1)}{\sqrt{P^{\mathsf{X}}(1)\,P^{\mathsf{X}}(0)\,P^{\mathsf{Y}}(1)\,P^{\mathsf{Y}}(0)}}$$

where $\mathbf{E}(\cdot)$ and $\mathbf{D}(\cdot)$ denote the mean and variance, respectively (for definitions of these notions refer to Casella and Berger [15]). Correlation of binary variables is undefined, if one or both of the variables are almost surely equal to 0 or almost surely equal to 1. Otherwise its value ranges from -1 (perfect anticorrelation) to $+1$ (perfect correlation). These extremes occur in a situation where almost surely $X = -Y$ or $X = Y$, respectively. Binary random variables are independent if and only if their correlation is 0.

The other two measures originate from epidemiological studies comparing the risk of some event happening in one or another group of individuals [41]:

Absolute risk change $$\mathsf{ARC}(Y|X) = \left|P^{\mathsf{Y}|\mathsf{X}}(1|1) - P^{\mathsf{Y}|\mathsf{X}}(1|0)\right|$$

Relative risk $$\mathsf{RR}(Y|X) = \frac{P^{\mathsf{Y}|\mathsf{X}}(1|1)}{P^{\mathsf{Y}|\mathsf{X}}(1|0)}$$

Absolute risk change is also known as absolute risk increase or reduction depending on the direction of change. Absolute risk change and relative risk are non-symmetric measures, *i.e.* swapping the two binary random variables changes the value. Each of the equalities $\mathsf{ARC}(Y|X)=0$ and $\mathsf{RR}(Y|X)=1$ is equivalent to the independence of X and Y, assuming $0 < P^{\mathsf{X}}(1) < 1$ and $0 < P^{\mathsf{Y}}(1) < 1$.

Sample and empirical probability

Many properties of random variables can be learned from observations of a *sample*. An *i.i.d. sample* of size n from the distribution of a random variable W is a list of n independent random variables $W_1, W_2, \ldots, W_n$, which are all distributed identically to W. For any set $\mathcal{S}$ we define $N_n^{\mathsf{W}}(\mathcal{S})$ as the random variable representing the number of elements in the sample which take on the value from set $\mathcal{S}$. If the set $\mathcal{S}$ has a single element, $\mathcal{S} = \{s\}$, then we omit the curly braces, $N_n^{\mathsf{W}}(s) = N_n^{\mathsf{W}}(\{s\})$.

The *empirical probability* of $X{=}x$ for a binary random variable X and the *empirical cumulative distribution function* of a real-valued random variable W

are defined as

$$P_n^{\mathsf{X}}(x) = \frac{N_n^{\mathsf{X}}(x)}{n} \qquad\qquad F_n^{\mathsf{W}}(w) = \frac{N_n^{\mathsf{W}}\big((-\infty, w]\big)}{n}$$

for any $x \in \{0, 1\}$ and $w \in \mathbb{R}$. According to the strong law of large numbers the empirical probability converges to the probability almost surely, that is with probability 1:

$$\lim_{n\to\infty} P_n^{\mathsf{X}}(x) = P^{\mathsf{X}}(x) \quad \text{a.s.}$$

An i.i.d. sample of size n from the joint distribution of random variables W and W' is a list of n independent random vectors $(W_i, W_i')_{i=1}^n$, where each vector is distributed identically to (W, W'). For any sets $\mathcal{S}, \mathcal{S}'$ we define $N_n^{\mathsf{WW}'}(\mathcal{S}, \mathcal{S}')$ as the random variable representing the number of pairs in the sample for which $W_i \in \mathcal{S}$ and $W_i' \in \mathcal{S}'$. Again, the curly braces are omitted if a set has only a single element.

For binary random variables X and Y the empirical probability of the event $\{X{=}x, Y{=}y\}$ and the *empirical conditional probability* of Y given $X{=}x$ are defined as

$$P_n^{\mathsf{XY}}(x, y) = \frac{N_n^{\mathsf{XY}}(x, y)}{n} \qquad P_n^{\mathsf{Y}|\mathsf{X}}(y|x) = \frac{N_n^{\mathsf{XY}}(x, y)}{N_n^{\mathsf{X}}(x)} \qquad \forall x, y \in \{0, 1\}$$

For a binary random variable X and a real-valued random variable W the *empirical conditional cumulative distribution function* of W given $X{=}x$ is defined as

$$F_n^{\mathsf{W}|\mathsf{X}}(w|x) = \frac{N_n^{\mathsf{XW}}\big(x, (-\infty, w]\big)}{N_n^{\mathsf{X}}(x)} \quad \forall x \in \{0, 1\}\ \forall w \in \mathbb{R}$$

Empirical measures of association

Empirical measures of association measure the association between two random variables using an i.i.d. sample from the joint distribution of these variables. Most measures of association can be converted to the corresponding empirical measure by replacing probabilities with empirical probabilities in the defining formula. However, some empirical measures cannot be obtained this way and are defined directly based on the sample [60].

In the dissertation we use an empirical measure of association between a binary random variable X and a real-valued random variable W defined as follows:

$$\mathsf{KS}_n(W|X) = \sup_{w\in\mathbb{R}} \Big| F_n^{\mathsf{W}|\mathsf{X}}(w|1) - F_n^{\mathsf{W}|\mathsf{X}}(w|0) \Big|$$

This measure uses the well-known Kolmogorov-Smirnov distance to quantify the difference of empirical distributions of W in the two subsamples corresponding to $X{=}1$ and $X{=}0$.

Hypothesis testing

Hypothesis testing is a statistical technique for deciding between two rivalling user-defined hypotheses – the *null hypothesis* and the *alternative hypothesis* – based on observed data. The decision depends on the *test statistic*, which is a user-defined function measuring the extremality of data with respect to the null hypothesis. The test statistic is used to calculate the *p-value*, defined as the probability of obtaining the value of the test statistic at least as extreme as the one that was actually observed, assuming that the null hypothesis is true. If the p-value is below the threshold called the *significance level* and denoted as α, then the null hypothesis is *rejected* and the alternative hypothesis is announced. Otherwise, the null hypothesis is *accepted*, meaning that there is not enough evidence to reject it.

Multiple testing correction

Hypothesis testing can result in two types of errors. False positive is the case where the null hypothesis is rejected, while it is actually true. False negative is the case where the null hypothesis is accepted, while it is actually false. The probability of rejecting a true null hypothesis in a single hypothesis test is less than or equal to the significance level.

When a large number of tests has to be performed then even if the probability of erroneously rejecting a null hypothesis (the significance level) in each test is small, the probability of making *at least one* such error out of many can still be very high. Consequently, special procedures (*multiple testing correction*) must be used in order to control the amount of false positives in this setting.

Bonferroni correction is the simplest method, which suggests to reduce the p-value threshold from the single test significance level α down to the multiple tests significance level α/t where t is the number of tests. It can be proved that the probability of rejecting at least one true null hypothesis out of t after Bonferroni correction does not exceed the original significance level α.

1.3 Notations

In the dissertation we use the following notations and conventions.

$\mathbf{P}(A)$	the probability of event A	
$\mathbf{P}(A,\, B)$	the probability of events A and B occurring simultaneously	
$\mathbf{P}(A\|B)$	the probability of event A conditional to the event B	
a.s.	almost surely, means that the event occurs with probability 1	
i.i.d.	independent and identically distributed	
α	significance level	
$\mathcal{B}(1,p)$	Bernoulli distribution with mean p	
$\mathcal{H}(n,k,m)$	hypergeometric distribution with n balls total, k drawn and m marked	
$\mathcal{N}(\mu,\sigma^2)$	normal distribution with mean μ and variance σ^2	
X, Y	binary random variables	
W	a real-valued random variable	
n	size of a random sample	
X_i, Y_i, W_i	random variables representing the i-th element in the sample	
$N_n^{\mathsf{W}}(\mathcal{S})$	the number of elements in the sample with $W_i \in \mathcal{S}$	
$N_n^{\mathsf{XW}}(x,\, \mathcal{S})$	the number of elements in the sample with $X_i{=}x$ and $W_i \in \mathcal{S}$	
$\operatorname{corr}(X,Y)$	Pearson product-moment correlation coefficient	
$P^{\mathsf{X}}(x)$	probability of $X{=}x$	$= \mathbf{P}(X{=}x)$
$P^{\mathsf{XY}}(x,\, y)$	joint probability of $X{=}x$ and $Y{=}y$	$= \mathbf{P}(X{=}x,\, Y{=}y)$
$P^{\mathsf{Y}\|\mathsf{X}}(y\|x)$	conditional probability of $Y{=}y$ given $X{=}x$	$= \mathbf{P}(Y{=}y\|X{=}x)$
$F^{\mathsf{W}}(w)$	cumulative distribution function	$= \mathbf{P}(W{\leq}w)$
$F^{\mathsf{W}\|\mathsf{X}}(w\|x)$	conditional cumulative distribution function	$= \mathbf{P}(W{\leq}w\|X{=}x)$
$\mathsf{ARC}(Y\|X)$	absolute risk change	$= \left\|P^{\mathsf{Y}\|\mathsf{X}}(1\|1) - P^{\mathsf{Y}\|\mathsf{X}}(1\|0)\right\|$
$\mathsf{RR}(Y\|X)$	relative risk	$= P^{\mathsf{Y}\|\mathsf{X}}(1\|1) \,/ P^{\mathsf{Y}\|\mathsf{X}}(1\|0)$
$P_n^{\mathsf{X}}(x)$	empirical probability of $X{=}x$	$= N_n^{\mathsf{X}}(x) \,/ n$
$P_n^{\mathsf{XY}}(x,\, y)$	empirical joint probability of $X{=}x$ and $Y{=}y$	$= N_n^{\mathsf{XY}}(x,\, y) \,/ n$
$P_n^{\mathsf{Y}\|\mathsf{X}}(y\|x)$	empirical conditional probability	$= N_n^{\mathsf{XY}}(x,\, y) \,/ N_n^{\mathsf{X}}(x)$
$F_n^{\mathsf{W}}(w)$	empirical cumulative distribution function	$= N_n^{\mathsf{W}}((-\infty, w]) \,/ n$
$F_n^{\mathsf{W}\|\mathsf{X}}(w\|x)$	empirical conditional cumulative distribution function	$= \dfrac{N_n^{\mathsf{XW}}(x,\, (-\infty, w])}{N_n^{\mathsf{X}}(x)}$
$\mathsf{KS}_n(W\|X)$	two-sample Kolmogorov-Smirnov distance	$= \sup_{w\in\mathbb{R}} \left\|F_n^{\mathsf{W}\|\mathsf{X}}(w\|1) - F_n^{\mathsf{W}\|\mathsf{X}}(w\|0)\right\|$

CHAPTER 2

STATISTICAL ENRICHMENT ANALYSIS

Statistical enrichment has been used in various different contexts in bioinformatics, *e.g.* functional enrichment [32], motif enrichment [45], and positional enrichment [51]. However, no unifying definition of enrichment has been given to our knowledge. In this chapter we give intuitive and formal definitions of enrichment, and see how these definitions work in functional enrichment analysis and regulatory enrichment analysis. It appears that in both of these cases enrichment measures the association between two properties of genes. Finally, we describe the Fisher's exact test and the Kolmogorov-Smirnov test, that are used for studying association in several enrichment analysis methods.

2.1 Definition of enrichment

Enrichment analysis applies to the situation where we have some prior expectation about a quantity that can be calculated from data. The calculated quantity can then be either larger than expected (*enrichment* or *over-representation*), smaller than expected (*depletion* or *under-representation*) or the same (*no enrichment, no depletion*). The expectation, which we also call *reference*, is based on our assumptions and prior knowledge about how the data were obtained and how the quantity was calculated.

We illustrate the concept of enrichment and the related issues with an example of a vehicle being tested for emission of pollutants at technical inspection. The data about the vehicle and its emissions are gathered through some experimental protocol. These data might include the vehicle's manufacturer, the model, production year, type of engine, and the measured emission of different pollutants. The exhaust of the vehicle is said to be enriched in pollutants, if the emissions are higher than expected. Depending on the reference the interpretation of enrichment can be different. If the expectation is based on the norm as stated in the law,

then the vehicle with emission enriched in pollutants is violating the regulations. Alternatively, if the expectation is based on the specification by the vehicle's manufacturer, then the vehicle with emission enriched in pollutants is out of order, but not necessarily violating the regulations.

As experimental data almost surely involve random fluctuations, then they rarely show exactly the expected value for the quantity. This must be taken into account by the reference. Hence, we present the reference as a probability distribution over possible values of the quantity – the subjective probabilities of having one or another value of the quantity in the data. Now we say that the quantity is enriched (or depleted) only if it is unprobable to obtain a value as high (or as low) according to the reference distribution. This is in agreement with the technical inspection example, where the emissions test is passed unless the quantity of pollutants is too high to be explained by the measurement error.

To formalize the notion of enrichment we first denote the data by d, the quantity calculated from the data by $q(d)$, and the random variable specifying the reference distribution by R_d.

Definition 2.1. *The data d are enriched in quantity $q(\cdot)$ with respect to the reference R_d and significance level α, if the following holds:*

$$\mathbf{P}\Big(R_d \geq q(d) \Big) < \alpha$$

where $\mathbf{P}(\cdot)$ denotes probability, R_d is a real-valued random variable and $q(d) \in \mathbb{R}$.

According to this definition, testing enrichment is essentially performing a statistical hypothesis test where the quantity $q(\cdot)$ is the test statistic and R_d specifies the distribution of the quantity under the null hypothesis of no enrichment. The probability $\mathbf{P}(R_d \geq q(d))$ is the p-value, that is the probability of obtaining a result at least as extreme as the observed value $q(d)$, assuming that the null hypothesis is true.

As generally in statistical hypothesis testing, we are often not only interested in whether or not there is any effect (enrichment), but also in how strong the effect (enrichment) is. *Statistical enrichment analysis* addresses one or both of the following questions:

- Is there any enrichment?
- How strong is the enrichment?

Sometimes the strength of enrichment is measured by the same p-value which is used to check if there is any enrichment, *i.e.* smaller p-value corresponds to

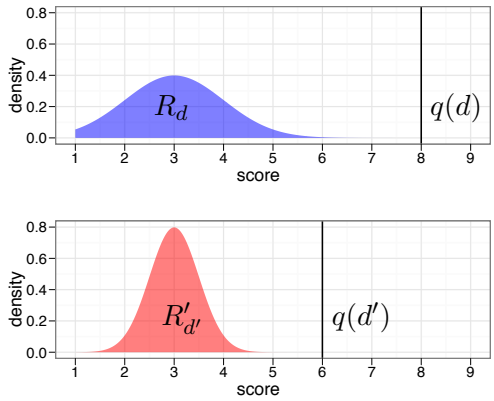


Figure 2.1: Enrichment of quantity $q(\cdot)$ in data d and d' with values $q(d) = 8$ and $q(d') = 6$ and references $R_d \sim \mathcal{N}(3, 1)$ and $R'_{d'} \sim \mathcal{N}(3, 0.5)$. The enrichment p-values are $\mathbf{P}(R_d \geq q(d)) \approx 3 \cdot 10^{-7}$ and $\mathbf{P}(R'_{d'} \geq q(d')) \approx 10^{-9}$. Different measures of enrichment strength can indicate stronger enrichment for one or the other case.

stronger enrichment. In the example presented in Figure 2.1, such criterion shows stronger enrichment for data d'.

However, suppose that the values $q(d)$ and $q(d')$ of this example are the results of measuring emission of the same pollutant in the exhaust of two different vehicles. In addition, suppose that the reference distributions represent the expected measurement results in the case where the true amount of the pollutant is 3, a hypothetical upper limit fixed by the law. The variance of the two reference distributions is smaller for d', suggesting usage of a more exact measurement device. With this interpretation of the example data in Figure 2.1, the vehicle corresponding to d seems to have stronger enrichment of the pollutant in the exhaust with respect to the allowed limit of 3, both visually and intuitively.

The example shows that the p-value from the definition of enrichment is not always appropriate for measuring the strength of enrichment. The choice of a suitable measure of enrichment strength depends very much on the context.

Statistical enrichment analysis can be viewed as a method of learning about any system, be it biological, ecological, physical or artificial. If the reference is chosen to correspond to our current understanding of the data, then the discovery or confirmation of enrichment can lead us to hypotheses about how to modify and improve our understanding.

2.2 Functional enrichment analysis

There are many software tools in bioinformatics using some kind of enrichment analysis. Huang *et al.* have published a systematic overview of 68 tools that perform *functional enrichment analysis* of gene lists [32]. Functional enrichment analysis is a family of methods which scan through many pre-compiled subsets of co-functioning genes and study how strongly these subsets are enriched in the genes input by the user. The reference for enrichment is usually determined by the situation where the user provides a random list of genes.

The pre-compiled subsets originate from databases covering many functional aspects of genes, like Gene Ontology [3] for biological processes, molecular functions or cellular components and KEGG [35] for pathways. The genes provided by the user are commonly determined by the results of high-throughput experiments. Functional enrichment indicates association between the experiments and function, guiding the biologists in further studies. For instance, if the user-given list is defined as the genes upregulated in some tumor compared to normal tissue, then the analysis might reveal association with the set of genes annotated to cell growth in the Gene Ontology database.

The simplest functional enrichment analysis tools (Class I according to Huang *et al.* [32]) input a subset of genes, compare it one by one to all pre-compiled subsets, and report those with unexpectedly high overlap with the input subset. Some of the tools (Class II) let the user add some real-valued score to each gene and find those pre-compiled subsets of genes which are enriched in the highest- or lowest-scoring genes.

A subset of genes can be treated as a binary property saying for each gene if it belongs to the subset or not. Therefore, the objective of these tools is to detect relations between a user-given binary (Class I) or real-valued (Class II) property and a set of pre-compiled binary properties. Most of the tools (all except Class III, for details see Huang *et al.* [32]) study the relation between two properties at a time, and decide whether the properties are independent (no enrichment) or associated (enrichment). Sometimes, the strength of association is also studied with various measures. The statistical methods for performing such analysis are discussed in Section 2.4.

2.3 Regulatory enrichment analysis

Another common application of enrichment analysis is in studying gene regulation. One of the major goals in this is to determine which DNA-binding transcription factors regulate the transcription of which genes. For many transcription factors the binding motifs are approximately known and gathered into databases

like Jaspar and Transfac [12, 44]. However, the motifs by themselves do not determine the regulation because binding can be affected by DNA methylation, nucleosomes and histone modifications [23].

Therefore, it is hard to decide based on the sequence alone whether a particular gene is regulated by a certain transcription factor. *Motif enrichment analysis* studies if the regulatory sequences of a given group of genes are enriched in known binding motifs of transcription factors. This way the sequence data from many genes are aggregated and enrichment indicates that the group of genes is regulated by the studied factor, without specifying which genes are and which not.

McLeay and Bailey [45] have presented a unifying framework of motif enrichment analysis highlighting two important decisions in the analysis. First, the framework requires the *motif affinity function* to be fixed. This function determines how each gene is scored for the presence of binding sites of the transcription factor. Second, the choice of the *association function* defines how to measure the association between the affinity scores and the given gene set. This is essentially measuring association between a real-valued and a binary property – a familiar task from functional enrichment analysis. The statistical methods for solving this task are discussed in the next Section 2.4.

All the motif affinity functions covered by McLeay and Bailey use only the genomic sequence information to score promoters for the presence of binding sites [45]. However, it has been shown that epigenetic information is essential for improving the binding site predictions [50]. Therefore, we propose an extension of this framework which can incorporate extra regulatory information.

The extension is straightforward – all regulatory information should be encoded as a single *regulatory scoring function*, which replaces the motif affinity function in the above framework. Consequently, the same association functions can be used to study the association between regulatory information and the given gene set. Thus, the new framework which we refer to as the *regulatory enrichment analysis*, is determined by the following two functions:

- **Regulatory scoring function** scores each gene for the potential to be regulated by the specified regulators;
- **Association function** detects association or measures the strength of the association between the regulatory score property and a given property of genes.

Note that this framework of regulatory enrichment analysis can in principle be applied to study any regulatory mechanisms of genes and is not restricted to study transcription factor binding. For example, one could test if the set of genes with

differential expression in some experiment is enriched in target genes of a microRNA [4].

2.4 Association between properties

In the previous two sections we have seen that functional and regulatory enrichment analyses require detecting and measuring the strength of the association between two properties of genes. The software tools which perform these analyses use many different statistical methods for this task [32, 45].

The existence of association between two binary properties is commonly tested with either Fisher's exact test, binomial test or chi-square test. If one of the properties is real-valued and the other binary, then the frequent choices are the Kolmogorov-Smirnov test, the Wilcoxon test, and the Student's t-test. Alternatively, some methods look for an optimal threshold to convert the real-valued property into binary, and by this reduce the task to testing the association of two binary properties [2, 55]. Note that in the context of functional and regulatory enrichment analyses these tests are usually repeated many times for different pairs of properties, and thus multiple testing correction is required.

Besides testing the existence of association, the functional and regulatory enrichment analyses often output information about the strength of association between the two properties of genes. In the simplest scenario, the strength is measured by the p-value resulting from the test of existence of association. Alternatively, any empirical measure of association can be used.

Next, we present the one-tailed Fisher's exact test and the two-sample Kolmogorov-Smirnov test as statistical enrichment tests which can detect the existence of association between two properties of genes or any other biological entities.

One-tailed Fisher's exact test

The one-tailed Fisher's exact test can be used for testing the existence of positive correlation between two binary properties. Let us denote the values of the two properties of gene i by x_i and y_i for $i = 1, \ldots, n$. In order to apply statistical methods we assume that the pairs (x_i, y_i) are observations of an i.i.d. sample $(X_i, Y_i)_{i=1}^n$ from the joint distribution of some binary random variables X and Y. Intuitively, we assume that the values of the properties of different entities are obtained independently and in an identical setting.

The variables X and Y are positively correlated, if

$$P^{\mathsf{XY}}(1,\,1) > P^{\mathsf{X}}(1)\,P^{\mathsf{Y}}(1)$$

One-tailed Fisher's exact test

Input data:	$D = (X_i, Y_i)_{i=1}^{n}$
Quantity:	$q(D) = N_n^{\mathsf{XY}}(1,\,1)$
Reference:	$R_D \sim \mathcal{H}(n, N_n^{\mathsf{X}}(1), N_n^{\mathsf{Y}}(1))$
Test:	$\mathbf{P}\Big(R_D \geq q(D) \Big) < \alpha$
Null hypothesis:	$\mathrm{corr}(X, Y) \leq 0$
Alternative hypothesis:	$\mathrm{corr}(X, Y) > 0$

Table 2.1: The one-tailed Fisher's exact test presented as an enrichment test which detects positive correlation between binary random variables X and Y based on an i.i.d. sample from the joint distribution of these variables. Here $\mathcal{H}(n, k, m)$ denotes the hypergeometric distribution and α is the significance level.

where we have used the notation introduced in Section 1.2 and summarized in Section 1.3. Estimating these probabilities empirically, we could check if

$$P_n^{\mathsf{XY}}(1,\,1) > P_n^{\mathsf{X}}(1)\, P_n^{\mathsf{Y}}(1)$$

or equivalently, if

$$N_n^{\mathsf{XY}}(1,\,1) > N_n^{\mathsf{X}}(1)\, N_n^{\mathsf{Y}}(1)\,/n$$

However, if the difference between the two sides of the latter inequality is small, then we do not know whether it is due to true correlation or just due to random fluctuations in the data. This can be decided using the one-tailed Fisher's exact test (also known as the *hypergeometric test*), which is presented as an enrichment test in Table 2.1. It tests whether the data are enriched in the quantity $N_n^{\mathsf{XY}}(1,\,1)$ with respect to the hypergeometric reference which expects independence of X and Y. Or in other words, it tests whether the number of genes having both properties equal to 1 at the same time is significantly higher than expected to be if the properties would be unrelated. Note that in Table 2.1 we have presented data as a random vector D instead of a fixed value d as we know that the data are a random sample.

The Fisher's exact test is applied in most of the Class I tools of functional enrichment analysis [32] and in several motif enrichment tools [27, 31, 43, 55].

Two-sample Kolmogorov-Smirnov test

The two-sample Kolmogorov-Smirnov test can be used for testing the existence of association between a binary and a real-valued property. Let us denote the values of the binary and real-valued property of gene i by x_i and w_i for $i = 1, \ldots, n$. As for the Fisher's exact test, we assume the pairs (x_i, w_i) to be observations of an

Two-sample Kolmogorov-Smirnov test

Input data:	$D = (X_i, W_i)_{i=1}^n$
Quantity:	$q(D) = \sqrt{\frac{N_n^{\mathsf{X}}(1) N_n^{\mathsf{X}}(0)}{N_n^{\mathsf{X}}(1) + N_n^{\mathsf{X}}(0)}} \cdot \sup_{w \in \mathbb{R}} \left\lvert F_n^{\mathsf{w}\mid\mathsf{x}}(w\mid 1) - F_n^{\mathsf{w}\mid\mathsf{x}}(w\mid 0) \right\rvert$
Reference:	$R_D \sim \mathcal{K}$
Test:	$\mathbf{P}\Big(R_D \geq q(D) \Big) < \alpha$
Null hypothesis:	X and W are independent
Alternative hypothesis:	X and W are associated

Table 2.2: The two-sample Kolmogorov-Smirnov test presented as an enrichment test which detects association between a binary random variable X and a real-valued random variable W based on an i.i.d. sample from the joint distribution of these variables. Here $\mathcal{K}$ denotes the limiting distribution of the Kolmogorov-Smirnov statistic [57].

i.i.d. sample $(X_i, W_i)_{i=1}^n$ from the joint distribution of random variables X and W.

The variables X and W are associated, if the conditional cumulative distribution functions $F^{\mathsf{w}\mid\mathsf{x}}(w\mid 1)$ and $F^{\mathsf{w}\mid\mathsf{x}}(w\mid 0)$ are different. This occurs if and only if the maximum difference of these functions over all possible arguments is positive,

$$\sup_{w \in \mathbb{R}} \left| F^{\mathsf{w}\mid\mathsf{x}}(w\mid 1) - F^{\mathsf{w}\mid\mathsf{x}}(w\mid 0) \right| > 0$$

Estimating the conditional cumulative distribution functions empirically, we could check if

$$\sup_{w \in \mathbb{R}} \left| F_n^{\mathsf{w}\mid\mathsf{x}}(w\mid 1) - F_n^{\mathsf{w}\mid\mathsf{x}}(w\mid 0) \right| > 0$$

However, if the empirical maximum difference is only slightly larger than zero, then we do not know whether it is due to true association or just due to random fluctuations in the data. This can be decided using the two-sample Kolmogorov-Smirnov test, which is presented as an enrichment test in Table 2.2. Note that the sample of size n is split into two subsamples based on the values of X_i, and thus

$$N_n^{\mathsf{x}}(1) + N_n^{\mathsf{x}}(0) = n$$

The Kolmogorov-Smirnov test is used in some Class II tools for functional enrichment analysis, such as GeneTrail [5], GOdist [7] and GSEA [59]. The test statistic $q(D)$ without the square root term is known as the Kolmogorov-Smirnov distance between samples and will also be used in Section 4.2.

CHAPTER 3

HIERARCHICAL CLUSTERING AND FUNCTIONAL ENRICHMENT

In the previous chapter we introduced different forms of statistical enrichment analysis, in the three remaining chapters we apply these for studying gene regulation. In the current chapter we introduce the Papers I and II, where we have proposed novel methods for fast approximate hierarchical clustering and hierarchical functional enrichment analysis of clustered gene expression data.

3.1 Motivation

One of the first and most important steps in studying gene regulation is to measure gene expression in different cell types, developmental stages, pathological states and environmental conditions. This is commonly done with either gene expression microarrays [56] or in recent years also with RNA-seq [48] based on next generation sequencing [42]. These are high-throughput technologies which can provide the expression levels of most of the genes in a sample simultaneously.

A typical second step is grouping the genes or conditions by similarity of expression [22, 52]. This allows for better visualization of the data as well as supports further analyses. Alternatively or complementarily to grouping, genes expressed differentially between several conditions are determined [36].

Hierarchical clustering is a technique used often for grouping the gene expression data [22]. It builds a hierarchy of groups or clusters, such that each cluster which has at least two entities consists of two smaller clusters [34]. The result can be depicted as a tree called *dendrogram*.

A standard method for performing hierarchical clustering is *agglomerative hierarchical clustering* with different alternatives for *linkage*, *i.e.* for choosing which clusters to merge at each step [34]. This method is highly configurable,

allowing any similarity measure to be used. The complexity of the algorithm depends on the linkage method but is at least quadratic in the number of clustered items, as all pairwise similarities have to be calculated [18]. For large gene expression data sets this can take several minutes or even hours (see details in Paper I).

3.2 Fast approximate hierarchical clustering (Paper I)

Paper I introduces the concept of approximate hierarchical clustering and proposes an algorithm called HappieClust for performing it fast. HappieClust is especially suited for interactive applications where users expect a fast response but at the same time are not willing to give up on quality.

The key to the algorithm is to limit the number of calculated pairwise distances to a carefully chosen subset of all possible pairs. For this we have developed a heuristic producing a subset of object pairs, which is enriched in pairs with smaller distances (empirical data shown in Figure 1 of Paper I on p.74). Knowing pairs of similar objects is of critical importance in mimicking the greedy choices of full hierarchical clustering.

The heuristic relies on the geometric properties of the data space, particularly the triangle inequality which states that the distance from A to B cannot be longer than the sum of distances from A to C and C to B. A direct corollary from this is that if A and B are very close to each other, then the distance to any C from A and B is approximately the same. The heuristic turns this observation upside down and looks for pairs of objects which are approximately at the same distance from several other objects which are referred to as *pivots*. Pivots are used widely in the methods for performing similarity search [66]. The proposed approximate hierarchical clustering algorithm HappieClust performs the following steps:

1. A small set of pivots is chosen randomly (*e.g.* 20 pivots).
2. The distance from each object to each pivot is calculated.
3. The heuristic is used to obtain a subset of pairs enriched in pairs of similar objects.
4. A random subset of pairs is added to the heuristical subset, a step experimentally shown to improve the quality of approximation (results shown in Figure 4 of Paper I on p.78).
5. Agglomerative hierarchical clustering is performed using an algorithm modified to work with a subset of all pairs of distances.

Besides the same inputs as full hierarchical clustering, HappieClust additionally requires the user to specify the number of pivots, the number of distances to be calculated, and the expected proportion between heuristical and random pairs. It is also possible to provide a limit for the program running time. In that case, HappieClust dynamically chooses the appropriate number of distances to calculate. Computational experiments show that 20 pivots and an equal number of heuristical and random pairs (*i.e.* proportion 0.5) are choices which work well most of the time (see Figure 4 of Paper I on p.78).

Finally, the suitability of approximate hierarchical clustering for gene expression data clustering is evaluated. For this three different strategies are used to measure the quality of a dendrogram.

- *Joining distance ratio* adds up the distances between all pairs of clusters that are merged at some point in HappieClust and compares this to the respective sum for full hierarchical clustering.
- *Subtree content conservation* studies how compactly the objects in one subtree of full hierarchical clustering are positioned in the approximate dendrogram.
- *Functional enrichment conservation* studies if the functional enrichment of the genes in some subtree of full clustering is preserved in the approximate clustering.

The analysis of the computational experiments reveals that with a large dataset most of the biologically meaningful clusters can be obtained more than an order of magnitude faster. With clusters of more than 200 genes Happieclust performed on full data almost as well as the full clustering on 90% of the data. This suggests that the approximation error of HappieClust can be almost as small as the natural variance in the data.

3.3 Hierarchical functional enrichment analysis of microarray data (Paper II)

Since the genes with similar function tend to be co-expressed, functional enrichment analysis can be used to provide biological interpretation for the clusters of gene expression data [3, 37]. This can also help the biologist to find interesting clusters from a dendrogram containing tens of thousands of genes. Paper II introduces a web server VisHiC for clustering and visualization of gene expression data combined with automated functional enrichment analysis.

VisHiC inputs a gene expression data set, performs hierarchical clustering with HappieClust, functional enrichment analysis with g:Profiler [54], and finally

visualizes the results. The results are represented in a similar manner as in many gene expression data visualization tools, where the dendrogram is accompanied with a heatmap specifying the color-coded expression levels of all genes in all conditions. The main difference of VisHiC is that it can provide a compact view where functionally most relevant clusters according to enrichment data are highlighted and summarized, whereas the remaining genes are hidden.

VisHiC provides two alternatives for measuring relevance of a cluster. First measure is just the best p-value from the functional enrichment analysis of the cluster, whereas the second adds up the p-values for all significant functional annotations after logarithmic transformation. The latter measure is divided by the size of the cluster as larger clusters tend to have more and better p-values. Summarization is performed according to the relevance measure starting from the most relevant ones and avoiding clusters which already have some subcluster summarized.

Once the visualization is generated, the web server allows to zoom into the summarized clusters to see the full data, functional enrichment information and a lineplot with the expression profiles of genes in the cluster.

CHAPTER 4

REGULATORY ENRICHMENT IN PROMOTER ANALYSIS

In this chapter we give a detailed description and a theoretical explanation of the methods used for performing promoter analysis in the Papers III and IV.

4.1 Motivation

In the previous chapter we studied clusters of co-expressed genes and discussed that these often share annotations about the molecular function, biological process and cellular component. In addition to that, the co-expressed genes often share the regulatory mechanisms [46]. Specifically, Meng *et al.* [46] have confirmed that in many mammalian transcription factor manipulation experiments the promoter sequences of co-expressed genes are enriched in binding motifs of the manipulated factor. Therefore, such regulatory enrichment in the co-expressed genes of some process can point to transcription factors which are important during this process. This has been used in several studies [24, 64].

Several bioinformatics software tools have been developed to discover such regulatory enrichment of transcription factor binding motifs, such as Toucan [1], Clover [26], oPOSSUM [31], PAP [16], CORE_TF [30], ASAP [43], Pscan [65], PASTAA [55], FactorY [27] and AME [45]. All these tools scan through a large set of known transcription factors and calculate some score allowing to prioritize the factors with respect to enrichment. These tools cover a wide variety of regulatory scoring and association functions. Most of the tools provide a threshold to decide whether the enrichment score is statistically significant.

Our final goal in this chapter is to perform such regulatory enrichment analysis on two biological cases. For this we propose a novel method of measuring association and therefore do not use any of the above-mentioned tools.

4.2 The proposed method for measuring association

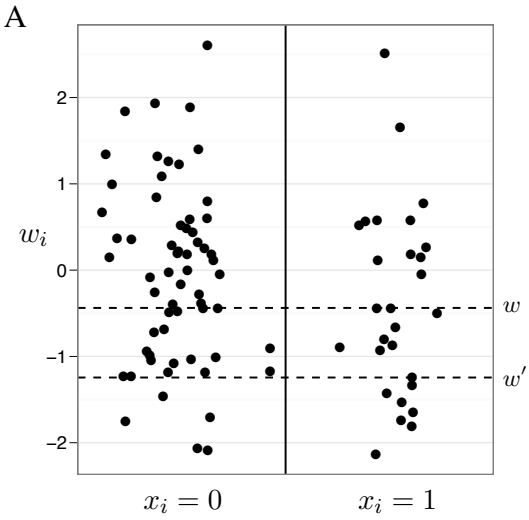
Here we propose a novel method for measuring the strength of association between a given subset of genes and their potential regulator. For each gene we use two pieces of information. First, a binary value x_i denotes whether the gene i belongs to the given subset ($1 =$ belongs, $0 =$ not). Second, a real value w_i denotes the regulatory score we have somehow obtained for this gene. In addition, we assume that there is a hidden binary value y_i denoting whether the gene i is actually regulated by the regulator under study ($1 =$ regulated, $0 =$ not). We are interested in quantifying the enrichment of the given subset in regulated genes.

As in Section 2.4 where we studied the association of properties, we assume here that vectors (x_i, y_i, w_i) are observations of an i.i.d. sample $(X_i, Y_i, W_i)_{i=1}^n$ from the joint probability distribution of some random variables X, Y and W. We further assume that X and W are independent conditional to Y, *i.e.* the regulatory scores W say nothing about the given set X in addition to what is already said by regulation Y. This is a reasonable assumption, if the experimental procedures for obtaining x_i and w_i are using different types of data, such as the expression and sequence data. Formally, we do not require anything else, but the proposed association measure will be effective only if smaller regulatory scores suggest higher probability of regulation. If the situation is the opposite, the negated scores $-w_i$ should be used instead of w_i.

Since X and Y are binary, the assumption of conditional independence can be formulated as the following statistical model for some $s, p, q \in [0, 1]$ and for some probability distributions $\mathcal{R}\mathsf{eg}$ and $\mathcal{N}\mathsf{eg}$:

$$\begin{aligned} X &\sim \mathcal{B}(1, s) \\ Y|X{=}1 &\sim \mathcal{B}(1, p) \\ Y|X{=}0 &\sim \mathcal{B}(1, q) \\ W|Y{=}1 &\sim \mathcal{R}\mathsf{eg} \\ W|Y{=}0 &\sim \mathcal{N}\mathsf{eg} \end{aligned} \tag{4.1}$$

where $\mathcal{B}(1, p)$ denotes the Bernoulli distribution with parameter p. The particular order $X \to Y \to W$ of conditioning was chosen because below we will use the values $p = P^{\mathsf{Y}|\mathsf{X}}(1|1)$ and $q = P^{\mathsf{Y}|\mathsf{X}}(1|0)$ to quantify the association between X and Y. The random variables W and X are indeed independent conditional to Y in this model, since the distribution of W depends on Y, but not on X directly. Note that there are no restrictions on the distributions $\mathcal{R}\mathsf{eg}$ and $\mathcal{N}\mathsf{eg}$, *i.e.* they do not have to belong to any known family of probability distributions. The parameters s, p, q and distributions $\mathcal{R}\mathsf{eg}$ and $\mathcal{N}\mathsf{eg}$ in the model (4.1) have the following



B

$s = 1/3$
$p = 0.60$
$q = 0.15$
$\mathcal{R}\mathsf{eg} = \mathcal{N}(-1, 1)$
$\mathcal{N}\mathsf{eg} = \mathcal{N}(0, 1)$

C

$\mathsf{ARC}(Y|X) = |p - q| = 0.35$
$\mathsf{KS}^*(W|X) \approx 0.13$
$\mathsf{KS}_n(W|X) = |\frac{16}{28} - \frac{22}{62}| \approx 0.22$
$w = -0.44$

$\mathsf{RR}(Y|X) = p/q = 4.00$
$\mathsf{Q}^*(W|X) = 4.00$
$\mathsf{Q}'_n(W|X) = \frac{8/28}{5/62} \approx 3.54$
$w' = -1.24$

Figure 4.1: Example data with $n = 90$ data points (w_i, x_i), where 62 points have $x_i = 0$ and 28 points have $x_i = 1$. **(A)** A plot of the data points, where the points are horizontally randomly positioned inside the panels $x_i = 1$ and $x_i = 0$ for better overview. **(B)** The parameters of the model (4.1) used to generate the above data points. **(C)** Non-empirical and empirical measures of association between the random variables. The optimal thresholds from the calculation of $\mathsf{KS}_n(W|X)$ and $\mathsf{Q}'_n(W|X)$ are denoted as w and w', respectively. The significance level in the calculation of $\mathsf{Q}'_n(W|X)$ was $\alpha = 0.05$.

interpretations:

s – the expected proportion of given genes among all genes;

p – the expected proportion of regulated genes among the given genes;

q – the expected proportion of regulated genes among the non-given genes;

$\mathcal{R}\mathsf{eg}$ – the expected distribution of scores among the regulated genes;

$\mathcal{N}\mathsf{eg}$ – the expected distribution of scores among the non-regulated genes.

Figure 4.1A plots example data with 90 data points drawn from the model (4.1) with parameters specified in Figure 4.1B.

Association between X and Y

Recall that our goal is to measure the strength of association between a given subset of genes and their potential regulator. In the current notation, this means the association between X and Y, whereas we know only the values $(x_i, w_i)_{i=1}^n$. The task might seem unsolvable, but due to the independence of X and W conditional to Y it is possible to say something about the association between X and Y.

We find p and q to be intuitive starting points for quantifying this association, as these measure the expected proportion of regulated genes among the given and non-given genes, respectively. Regulatory enrichment of the given subset of genes would indicate more regulated genes among the given genes, $p > q$. Two natural association measures based on p and q are the absolute risk change and the relative risk,

$$\mathsf{ARC}(Y|X) = \left| P^{\mathsf{Y}|\mathsf{X}}(1|1) - P^{\mathsf{Y}|\mathsf{X}}(1|0) \right| = |p - q|$$
$$\mathsf{RR}(Y|X) = \frac{P^{\mathsf{Y}|\mathsf{X}}(1|1)}{P^{\mathsf{Y}|\mathsf{X}}(1|0)} = \frac{p}{q}$$

representing the difference and fold-change of the expected proportion of regulated genes between given and non-given genes. According to the absolute risk change the association between X and Y with $p = 0.59$ and $q = 0.50$ is the same as with $p = 0.10$ and $q = 0.01$, since the difference is 0.09 in both cases. We decided the latter case to be biologically more interesting, and therefore preferred relative risk for the promoter analyses in the Papers III and IV. However, in the following theoretical subsection we consider both association measures.

Empirical association between X and W is an approximate lower bound for association between X and Y

As the values y_i are hidden, we cannot measure the association between X and Y directly. In the following we prove that the two-sample Kolmogorov-Smirnov distance $\mathsf{KS}_n(W|X)$ and a proposed empirical association measure $\mathsf{Q}_n(W|X)$ are approximate lower bounds for $\mathsf{ARC}(Y|X)$ and $\mathsf{RR}(Y|X)$, respectively.

The empirical measures $\mathsf{KS}_n(W|X)$ and $\mathsf{Q}_n(W|X)$ both involve testing different thresholds $w \in \mathbb{R}$ and calculating the proportion of data points with $w_i \leq w$ among the points with $x_i = 1$ and among the points with $x_i = 0$. For any w, these proportions are the values of the empirical conditional cumulative distribution functions $F_n^{\mathsf{W}|\mathsf{X}}(w|1)$ and $F_n^{\mathsf{W}|\mathsf{X}}(w|0)$, respectively (for statistical preliminaries see Section 1.2). For example, in Figure 4.1A the value $F_n^{\mathsf{W}|\mathsf{X}}(w'|1)$ is $8/28$, because there are 28 data points with $x_i = 1$ and 8 of those have $w_i \leq w'$. The measure $\mathsf{KS}_n(W|X)$ is defined as the maximum difference between $F_n^{\mathsf{W}|\mathsf{X}}(w|1)$ and $F_n^{\mathsf{W}|\mathsf{X}}(w|0)$ over all thresholds $w \in \mathbb{R}$,

$$\mathsf{KS}_n(W|X) = \sup_{w \in \mathbb{R}} \left| F_n^{\mathsf{W}|\mathsf{X}}(w|1) - F_n^{\mathsf{W}|\mathsf{X}}(w|0) \right| \tag{4.2}$$

The value of this measure and the maximizing threshold w for the example data are given in Figure 4.1C. Although the maximum is taken over all $w \in \mathbb{R}$, it is

enough to find the maximum over all values $w = w_i$, because the value of functions $F_n^{\mathsf{W}|\mathsf{x}}(w|1)$ and $F_n^{\mathsf{W}|\mathsf{x}}(w|0)$ changes only when w passes the data points. The measure $\mathsf{KS}_n(W|X)$ is the Kolmogorov-Smirnov distance between empirical distributions of W in the two subsamples corresponding to $X=1$ and $X=0$. Theorem 4.1 proves that with sample size n growing to infinity, $\mathsf{KS}_n(W|X)$ converges almost surely to a value $\mathsf{KS}^*(W|X)$, which is a lower bound for $\mathsf{ARC}(Y|X)$.

Theorem 4.1. *Let X, Y, W be random variables satisfying the statistical model (4.1) where $0 < s < 1$. If $(X_i, W_i)_{i=1}^n$ is an i.i.d. sample from the joint distribution of X and W, then the following holds:*

$$\lim_{n\to\infty} \mathsf{KS}_n(W|X) \overset{\text{a.s.}}{=} \mathsf{KS}^*(W|X) \leq \mathsf{ARC}(Y|X)$$

*where $\mathsf{KS}_n(W|X)$ is defined by the equality (4.2), the convergence occurs almost surely (*a.s.*) and*

$$\mathsf{KS}^*(W|X) = \sup_{w\in\mathbb{R}} \left| F^{\mathsf{W}|\mathsf{x}}(w|1) - F^{\mathsf{W}|\mathsf{x}}(w|0) \right|$$

Proof. See Appendix A. □

The intuition behind Theorem 4.1 is that as X and W are independent conditional to Y, then any association between these must be due to the associations between X and Y and between Y and W.

The measure $\mathsf{Q}_n(W|X)$ is defined using the same functions $F_n^{\mathsf{W}|\mathsf{x}}(w|1)$ and $F_n^{\mathsf{W}|\mathsf{x}}(w|0)$ as $\mathsf{KS}_n(W|X)$, but it maximizes the ratio of these instead of the difference. As for small values of w the value of $F_n^{\mathsf{W}|\mathsf{x}}(w|0)$ becomes small and thus the variance of the ratio becomes huge, we restrict w to larger values, specified by the set $\mathcal{B}_n$:

$$\mathsf{Q}_n(W|X) = \sup_{w\in\mathcal{B}_n} \frac{F_n^{\mathsf{W}|\mathsf{x}}(w|1)}{F_n^{\mathsf{W}|\mathsf{x}}(w|0)} \quad \mathcal{B}_n = \left\{ w \in \mathbb{R} \,\middle|\, F_n^{\mathsf{W}|\mathsf{x}}(w|0) > n^{-\beta} \right\} \tag{4.3}$$

The restrictive set $\mathcal{B}_n$ and the measure $\mathsf{Q}_n(W|X)$ are parametrized by β, where $0 < \beta < 1/4$. Such range of values for β is chosen because according to the Theorem 4.2 the quantity $\mathsf{Q}_n(W|X)$ converges for any such β almost surely to a value $\mathsf{Q}^*(W|X)$, which is a lower bound for $\mathsf{RR}(Y|X)$.

Theorem 4.2. *Let X, Y, W be random variables satisfying the statistical model (4.1) where $0 < s < 1$ and $0 < q < p < 1$. If $(X_i, W_i)_{i=1}^n$ is an i.i.d. sample from the joint distribution of X and W, then the following holds for any $0 < \beta < 1/4$:*

$$\lim_{n\to\infty} \mathsf{Q}_n(W|X) \overset{\text{a.s.}}{=} \mathsf{Q}^*(W|X) \leq \mathsf{RR}(Y|X)$$

where $\mathsf{Q}_n(W|X)$ *is defined by the equalities (4.3), the convergence occurs almost surely (*a.s.*) and*

$$\mathsf{Q}^*(W|X) = \sup_{w\in\mathbb{R}} \frac{F^{\mathsf{W}|\mathsf{X}}(w|1)}{F^{\mathsf{W}|\mathsf{X}}(w|0)}$$

Proof. See Appendix A. □

The empirical association measure used in the Papers III and IV

Note that the Theorems 4.1 and 4.2 describe only the asymptotic behaviour of $\mathsf{KS}_n(W|X)$ and $\mathsf{Q}_n(W|X)$. The theorems do not specify the quality of these estimates for any fixed n. In particular, Theorem 4.2 does not provide the best choice of β for specifying the restrictive set $\mathcal{B}_n$ for a fixed n. Therefore, further theory or computational experiments are required before using these methods.

As mentioned before, we decided to prefer relative risk to absolute risk change in the promoter analysis, as it coincided better with our understanding of what is biologically interesting. But since the Papers III and IV were published before we obtained the statement and proof of Theorem 4.2, we approximated relative risk with a modification of measure $\mathsf{Q}_n(W|X)$. The used measure $\mathsf{Q}'_n(W|X)$ differs from $\mathsf{Q}_n(W|X)$ in the restriction on w, and is defined as follows:

$$\begin{aligned}
\mathsf{Q}'_n(W|X) &= \sup_{w\in\mathcal{B}'_n} \frac{F_n^{\mathsf{W}|\mathsf{X}}(w|1)}{F_n^{\mathsf{W}|\mathsf{X}}(w|0)} \\
\mathcal{B}'_n &= \left\{ w\in\mathbb{R} \,\middle|\, \mathbf{P}\Big(R_D \geq N_n^{\mathsf{XW}}(1,(-\infty,w]) \Big) < \alpha \right\} \qquad (4.4)\\
R_D &\sim \mathcal{H}(n, N_n^{\mathsf{X}}(1), N_n^{\mathsf{W}}((-\infty,w]))
\end{aligned}$$

where α is the significance threshold and $\mathcal{H}(n,k,m)$ is the hypergeometric distribution. The definition of $\mathcal{B}'_n$ is based on the Fisher's exact test (see Table 2.1) for testing if the two binary random variables X and $W\leq w$ are associated. We cannot claim significant association as the same test is performed for many values of w and there is no multiple testing correction. But since the binary random variables $W\leq w$ for different w are highly correlated, then the set $\mathcal{B}'_n$ is non-empty only if there is quite some evidence for association.

Similar search for an optimal threshold with multiple Fisher's exact tests has been used earlier, but with different objective functions [2, 55].

4.3 Computational experiments

In this and the following sections of this chapter we use the short notation of Q'_n and RR instead of $\mathsf{Q}'_n(W|X)$ and $\mathsf{RR}(Y|X)$. In order to test how often

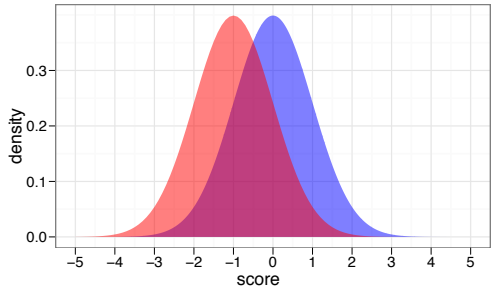


Figure 4.2: Distributions $\mathcal{R}\text{eg} = \mathcal{N}(-1, 1)$ and $\mathcal{N}\text{eg} = \mathcal{N}(0, 1)$ that were used to generate artificial data for computational experiments.

Q'_n is a lower bound for RR, we carried out some computational experiments. The data were generated as an i.i.d. sample from the statistical model (4.1) with $\mathcal{R}\text{eg} = \mathcal{N}(-1, 1)$ and $\mathcal{N}\text{eg} = \mathcal{N}(0, 1)$. Figure 4.1C presents the first example where Q'_n is a lower bound for RR.

For large scale experimentation we decided to test 100 and 1000 as the expected numbers of given and non-given genes. This fits approximately the order of magnitude of data in Papers III and IV. Hence, we chose sample size $n = 1100$ and $s = 1/11$. For p and q we tested all combinations of values $0.01, 0.02, \ldots, 0.99$. The distribution $\mathcal{R}\text{eg}$ was chosen to be shifted towards smaller values compared to $\mathcal{N}\text{eg}$ because the measure Q'_n becomes useful ($\mathsf{Q}'_n > 1$) only if $F_n^{\mathsf{W}|\mathsf{X}}(w|1)$ is greater than $F_n^{\mathsf{W}|\mathsf{X}}(w|0)$ for some w. The chosen shift was small to make it hard enough to distinguish between regulated and non-regulated genes based on the regulatory score (see Figure 4.2).

Figure 4.3A presents the probabilities $\mathbf{P}(\mathsf{Q}'_n \geq 1.8)$ estimated in the experiments for all combinations of values p and q. Here we consider the particular value 1.8 because in Paper IV we reported only the cases with $\mathsf{Q}'_n \geq 1.8$ to limit the number of results. The figure shows that the rate of false positives is very low, *i.e.* the cases with relative risk below 1.8 are almost never reported. This adds confidence to the relevance of the reported cases.

Theorem 4.2 proved that $\mathsf{Q}_n(W|X)$ is an asymptotic lower bound for relative risk. Figure 4.3B presents the empirical conditional probabilities that a reported Q'_n is indeed a lower bound for relative risk. As Figure 4.3A says that $\mathsf{Q}'_n \geq 1.8$ implies $\mathsf{RR} \geq 1.8$ with high probability, then by combining the Figures 4.3A and 4.3B we can say that $\mathsf{Q}'_n \geq 1.8$ implies $\mathsf{RR} \geq \mathsf{Q}'_n \geq 1.8$ with high probability. This follows from the fact that the region with high reporting probability in Figure 4.3A shows high probability of being a lower bound in Figure 4.3B. If all combinations of values of p and q would be equally likely, then the probability of $\mathsf{Q}'_n \leq \mathsf{RR}$

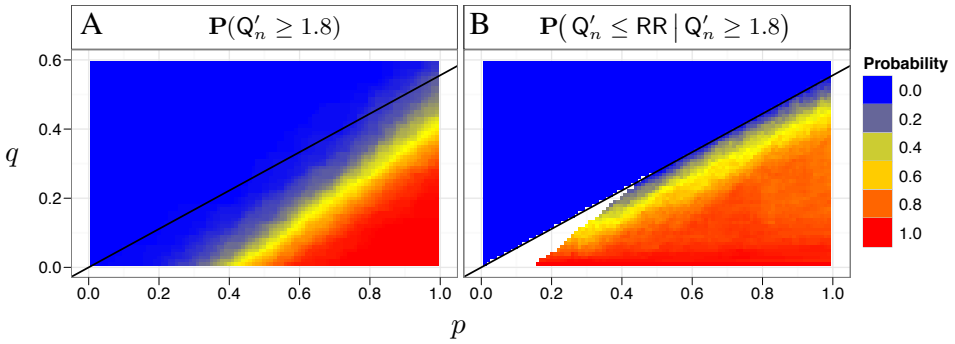


Figure 4.3: The results of computational experiments for different values of p and q. The black line refers to the points with $\mathsf{RR} = p/q = 1.8$. The probabilities were estimated from 100 repeated experiments in (A) and from experiments with 100 reported cases in (B). The white area indicates the region for which not enough experiments were performed as the reporting probability was very low. **(A)** The probability that the case was reported. **(B)** The conditional probability that Q'_n was a lower bound of RR, if it was known to be reported.

given $\mathsf{Q}'_n \geq 1.8$ would be about 0.8.

To conclude, the experiments have indicated that for a realistic value $n = 1100$ our association measure Q'_n is a lower bound of relative risk with high probability. We now proceed to the biological studies in Papers III and IV where we have used the measure Q'_n.

4.4 Promoter analysis of genes differentially expressed in mouse mesoderm development (in Paper III)

Paper III studies early development of mesoderm, the part of the embryo which develops into cell types such as muscles, heart, blood and kidney. Brachyury (also known as T) is a gene which has earlier been shown to be important in the development of mesoderm. Paper III applied a treatment technique on a transgenic mouse embryonic stem cell line to obtain brachyury$^+$ cells – 6-day-old embryoid bodies enriched in brachyury expressing cells. Next, the transcripts up- or down-regulated in the brachyury$^+$ cells compared to embryonic stem cells and control embryoid bodies were determined using gene expression microarrays. Functional enrichment analysis was performed on the list of differentially expressed genes, showing associations with Gene Ontology [3] terms *blood vessel morphogenesis*, *placenta development* and *cell death*, and KEGG [35] pathways for MAPK and TGFβ signalling, for example.

Two transcripts (Larp2 and Ankrd34b) with function unknown and the least

amount of description in earlier literature were chosen for further investigation from the transcripts up-regulated specifically in the brachyury$^+$ cells. Larp2 and Ankrd34b were silenced using siRNA knockdown protocol in the embryonic stem cells which were then monitored for any morphological changes. Also, the relative mRNA expression levels of 10 genes were analyzed in these cells using qRT-PCR technology. As a result, Larp2 was hypothesized to be positively involved in regulation of BMP-2 expression. Ankrd34b was suggested to be a positive regulator of the ectoderm-dependent neurogenesis and a negative regulator of the mesoderm-dependent adipogenesis and hematopoiesis.

The task for the author of this dissertation was to perform promoter analysis of the transcripts upregulated exclusively in the brachyury$^+$ cells (105 genes). Genes up-regulated exclusively in the BMP-2$^+$ cells (266 genes) and exclusively in the α-MHC$^+$ cells (584 genes) were used as background. In the notations of the previous sections, we had $n = 105{+}266{+}584 = 955$ genes with size of the given subset of genes equal to 105. Our first step was regulatory scoring of genes, *i.e.* estimating the potential of each transcription factor to regulate each gene.

Transcription of a gene can be influenced by both, near and far events of transcription factor binding to the genome. We considered only the closeby region of 2,000 bp upstream of the transcription start site, which should cover the promoter region in most cases. In these regions we scanned for putative transcription factor binding sites by matching the known motifs of transcription factor binding preferences obtained from the databases Transfac [44] and Jaspar [12] as position weight matrices. The regulatory score of a motif for each gene was calculated as the average score of the three highest scoring matches in the promoter.

In the previous section we noted that Q'_n becomes useful if the regulatory scores corresponding to regulated genes are shifted towards smaller values, but here, on the contrary, the regulated genes have probably higher average scores. Therefore, during the calculation of Q'_n we counted the proportions of points with $w \geq w_i$ instead of $w \leq w_i$. It can also be interpreted as negating all the regulatory scores and only then calculating Q'_n to obtain the enrichment score.

As a result, two motifs showed values $\mathsf{Q}'_n > 1$. The enrichment scores for the motifs PPARα and ISRE were $\mathsf{Q}'_n = 5.2$ and $\mathsf{Q}'_n = 2.7$, respectively. The promoters of genes scoring higher than w which reached the maximum in Q'_n are visualized in Figure 3 of Paper III on p.105. The figure also illustrates how a gene can have a high regulatory score (calculated as the average score of the three best matches of a motif) with just a single strong match or alternatively, with two or three weaker matches. PPARα was earlier known to be involved in cell differentiation, in particular cardiomyocytes differentiation of mouse embryonic stem cells. ISRE was known to induce apoptosis, which is in accordance with the observation that genes annotated to cell death are enriched among the genes upregulated in the brachyury$^+$ cells.

4.5 Promoter analysis of genes differentially expressed in mouse adipogenesis (in Paper IV)

Paper IV studies adipogenesis, which is the development of fat cells, also known as adipocytes. Several treatment techniques were used to induce and inhibit differentiation of mouse embryonic stem cells towards the adipocyte lineage. Large scale gene expression profiling was performed to reveal differential expression between the inducing and inhibiting treatments at several time points. As a result, four clusters of genes were obtained as upregulated and downregulated genes in day 6 and day 11. In addition, the fifth cluster included genes which were differentially expressed (either up- or downregulated) on both days.

Functional enrichment analysis was performed for each of the five clusters using g:Profiler [54]. The results are provided in Figure 2 of Paper IV on p.120. The biological highlights of this enrichment analysis include the association of adipocyte development with blood vessel development, neural development and the Wnt pathway.

The remaining part of Paper IV studies the transcriptional control of adipocyte development. First, Figure 5 lists the transcription factors within the five clusters of genes which were differentially expressed. These provide a simple hypothesis set of factors regulating the adipocyte development. It was shown experimentally that 7 out of 11 tested transcription factors of clusters 1 and 3 (genes upregulated on days 6 and 11, respectively) were upregulated in stromal vascular fraction of white adipose tissue in young mice, compared to the adipocyte fraction. This adds confidence to the hypothesis that these factors regulate adipogenesis, because the stromal vascular fraction contains adipocyte progenitors, whereas the adipocyte fraction encompasses only mature adipocytes.

Finally, the task for the author of this dissertation was to perform in silico analysis of transcription factor binding site enrichment in the promoters of genes in clusters 1–5 (the given subsets of genes) with 175, 25, 126, 52, and 15 genes, respectively. The total number of genes in the study was $n = 20694$. The analysis was similar to the analysis for Paper III presented in the previous section, and we will therefore highlight only the differences.

First, we extended the 2,000 bp upstream region 1,000 bp downstream of the transcription start site. Second, as the transcription factor binding sites tend to be more evolutionarily conserved than the surrounding sequence, we also took into account the conservation rate of the putative binding sites in rodents and primates (*Euarchontoglires*). So in addition to the regulatory score used in Paper III (average score of three best matches) we used regulatory scores defined by the score of the single highest scoring match which is above a conservation threshold. We tested thresholds 0.7, 0.8, 0.9 and 1.0, where the latter stands for 100% conser-

vation. Enrichment score was obtained using the reversed Q'_n as described in the previous section, but with reporting threshold $\mathsf{Q}'_n \geq 1.8$.

There were no results for the small clusters 2, 4 and 5, but for clusters 1 and 3 we found respectively 16 and 14 motifs with $\mathsf{Q}'_n \geq 1.8$. The results are presented in Figure 7 of Paper IV on p.127, where we list for each motif the following values:

- *Conservation* — the evolutionary conservation threshold used for defining the regulatory scores;
- *Targets in cluster* — the number of genes in the cluster which had a regulatory score above the optimal threshold w from the calculation of Q'_n;
- *Enrichment ratio* — the value Q'_n;
- *Enrichment P-value* — the p-value of the Fisher's exact test applied for measuring the association between binary random variables X and $W \geq w$, where w is the optimal threshold from the calculation of Q'_n. Note that this p-value was used in deciding if the particular value w belongs to the restrictive set $\mathcal{B}'_n$, see formula (4.4).

Several of the motifs reported for clusters 1 and 3 have been associated with adipogenesis earlier, see details in Paper IV. In addition, the motifs CART1, PRRX2, and MEIS1 that were enriched in the promoters of genes upregulated on day 11 (cluster 3), were corresponding to transcription factors upregulated on day 6 (marked with stars in Figure 7 of Paper IV on p.127). Together with information from protein-protein interactions (see details in Paper IV) this supports the hypothesis that these transcription factors are important for adipogenesis.

CHAPTER 5

EVOLUTIONARY MODELS OF ENRICHMENT (PAPER V)

In this final chapter we propose an evolutionary model of regulatory sequences. It models the DNA substring distribution, which determines the abundance of all motifs, and is therefore the key to the enrichment of motifs.

5.1 Motivation

In the previous chapter we performed regulatory enrichment analysis, which used the known transcription factor binding motifs to suggest which transcription factors regulate a user-given set of genes. Methods of *de novo motif discovery* search novel motifs that are enriched in the regulatory sequences of the user-given genes. The found motifs are good candidates for being binding motifs for some transcription factors with yet uncharacterized binding preference. The reference for enrichment determining the expected abundance of the motif is specified by either a genomic region which has presumably less binding sites of transcription factors (called a *background sequence*), or a probabilistic model which is often learned from such region (called a *background model*) [17].

The simplest background generating model is the zero-order hidden markov model (HMM) [6, 17]. It specifies the probabilities for nucleotides and the nucleotide at each position is generated independently from others. It has been shown that higher-order models which specify the distribution of longer substrings (*e.g.* of length up to 5, *i.e.* fourth-order HMM) can improve the sensitivity of motif discovery [62]. However, sometimes even the higher-order models do not describe enough statistical properties of a true biological sequence, so some methods use a biological background sequence [53].

The choice of a good background is important, otherwise the performance decreases significantly [62]. There are at least two reasons for that. First, if the

motifs we would like to find have high abundance in the chosen set of background sequences, then the enrichment can be statistically too weak for detection. Second, any features of the studied sequences lacking in the background can create a plethora of enriched motifs which overwhelm the results and hide the relevant signal. For instance, if the studied sequences include relatively more adenines ("A" nucleotides) than the background, then we may see enrichment of adenine-rich motifs. Taking into account the effect of such differences is non-trivial [11].

There are several reasons why the distribution of substrings can be different in two regions of the genome. First, there can be different functionally important features which do not accept mutations at various sites of the sequence. Second, the rates of different types of mutations can in principle be different at various regions of the genome. For instance, the mutations from CpG to TpG occur at different rates depending on whether or not the cytosine is methylated [14]. Finally, the distributions can be different due to larger genomic rearrangement events, such as duplication, insertion or deletion of long sequences.

In Paper V we propose an evolutionary model of DNA substring distribution. Given the functionally important features and mutation rates in a sequence of length n, the model provides the expected distribution of substrings of length $k \leq n$. Such evolutionary processes are hard to capture with an HMM. For example, our model allows to change the mutation rate from CpG to TpG by changing one parameter, while it is not at all clear how to change a background modelling HMM to achieve the same effect. Our evolutionary model also supports incremental motif discovery which looks for new motifs and complements the background model with the discovered motifs iteratively.

5.2 The proposed evolutionary model

We chose to model the evolution of a regulatory genomic sequence of individuals in a population with asexual reproduction, meaning that each descendant has a single parent. In our model, the presence and multitude of functionally important features in the regulatory sequence determines the fitness of an individual, which is a quantitative property describing the reproductive capability. In addition, there can be mutations in the sequence during reproduction.

We assume that over time the fitness function and mutation probabilities remain the same. Therefore, we might expect that over long enough time a large enough population gradually stabilizes to have a specific proportion of each of the possible sequences. We call such state an *equilibrium*, *i.e.* the state where the expected proportion of individuals with any particular sequence in the next generation is the same as in the current generation. Theorem 1 in Paper V on p.137 proves that if fitness is positive for all sequences and the probability of any se-

quence mutating into any other is positive, then there exists a unique equilibrium distribution which can be calculated as the unique positive eigenvector of a matrix specified in the paper. Theorem 2 adds that the equilibrium is asymptotically reached from any starting distribution of sequences.

The remaining part of Paper V studies the possibilities of calculating the distribution of substrings of length k in an equilibrium distribution of sequences of length n. Determining the equilibrium distribution of full sequences is generally intractable as this would require solving an eigenproblem of dimension 4^n, a task too hard already for small values of n, such as $n = 20$. So the question is, how well can the substring distribution be approximated without finding the full equilibrium? In the paper we have suggested one possible approximation which requires solving the same kind of eigenproblem as for finding the equilibrium, but with dimension 4^k.

To check the applicability of this approximation we performed experiments to compare the approximated distribution with the exact substring distribution under equilibrium. As finding the exact equilibrium is computationally so expensive, we used a 2-letter-alphabet and sequence length $n = 8$. The mutation probability at each position was identical and independent from other positions. The fitness of a sequence was defined based on the number of occurrences of a specific substring in the sequence. To estimate the quality of approximation, we measured the Kullback-Leibler divergence per position from the exact distribution to the approximated distribution, and the Pearson correlation between these distributions.

The experiments were performed for various mutation probabilities and fitness functions. The results turned out to depend mainly on the point-mutation probability. Figures 1 and 2 in Paper V on p.142 show that with mutation rate 0.1 or higher the approximation of the substring distribution was very close to the exact distribution. While correlation remained moderately good for lower mutation rates also, the Kullback-Leibler divergence showed decreasing quality of approximation. According to Figure 2 in Paper V on p.142 this was apparently caused by substrings with moderate true frequency but very low approximated frequency. In this example it seems (data not shown) that if the mutation rate approaches zero then the exact equilibrium distribution converges to the uniform distribution over sequences with maximum fitness, not captured well by the approximation.

5.3 Discussion

The conducted experiments show that we were able to approximate the substring distribution generated by the evolutionary model fairly well in various conditions. However, too few conditions were tested to suggest any requirements that would guarantee high quality approximation. Further experiments should be performed

with the 4-letter-alphabet and a larger variety of fitness and mutation functions.

Currently we have studied how to calculate the substring distribution once we have fixed the model. It would be useful to be able to learn the parameters of the model from a given substring distribution. This could be applied for studying the evolution of a particular sequence. More generally, it would be interesting to know what information it is possible to extract from a substring distribution using some prior information and assuming genetic equilibrium.

The assumption of genetic equilibrium of a long sequence is strong and probably does not hold in a real-life population. Small population size, large genomic rearrangement events, horizontal gene transfer, and time-varying fitness and mutation functions – all these can work against equilibrium. However, even if the population is not in equilibrium with respect to the long sequence, the substring distribution might still be similar to the case of equilibrium, a hypothesis worth studying further.

CONCLUSIONS

The goal of this dissertation was to enhance and apply algorithms involving or related to statistical enrichment analysis for studying gene regulation. We have first provided a formal definition of enrichment and studied how functional and motif enrichment analyses fit into this definition. We revealed that many algorithms that apply these analyses internally study the association between two properties of genes.

Hierarchical clustering of gene expression data is often one of the first steps in studying gene regulation. We developed a new algorithm for performing fast approximate hierarchical clustering using similarity heuristics, suitable for interactive applications which require fast algorithms. To highlight functionally more relevant parts in the results of clustering gene expression data hierarchically, we developed a software tool which performs hierarchical functional enrichment analysis.

Motif enrichment analysis can be applied to reveal which transcription factors are potentially important in the regulation of a group of genes. In order to take into account any relevant data in addition to the genomic sequence, such as evolutionary conservation or DNA methylation or histone modification, we have proposed a framework of regulatory enrichment analysis. This analysis requires quantifying the differences of the given group of genes and the group of all other genes with respect to potential of being regulated by the factor under study. This can be done with the Kolmogorov-Smirnov distance, measuring the difference of the proportion of the potentially regulated genes in the two groups. We have also proposed a novel measure for this purpose, using fold-change instead of the difference between the proportions. We have proved that the Kolmogorov-Smirnov distance and our measure are respectively the approximate lower bounds for absolute risk change and relative risk of regulation associated with the given group of genes. We have applied the novel measure to perform regulatory enrichment analysis of gene promoter regions in two studies of mouse embryonic stem cells.

Motif discovery can be applied to find unknown regulatory motifs in the genome. It requires as input a sequence which is presumably enriched in such motifs compared to some background. The background is commonly either some ge-

nomic sequence or a hidden markov model learned from it. As the final contribution of the dissertation we proposed an evolutionary model of DNA substring distribution, which can be used as a background model in motif discovery. The model supports incremental motif discovery which looks for new motifs and complements the background model with the discovered motifs iteratively.

Hopefully we have convinced the reader in the wide applicability of statistical enrichment analysis in bioinformatics. Most commonly, it is used to reveal association between properties of genes or any other biological entities. The results should always be interpreted with care, as there can be various reasons for enrichment, starting from causal relationships anywhere in the data and ending with a bias caused by technical problems in the experimental protocol. Also, in general the lack of enrichment does not rule out the possibility of associations in the data.

APPENDIX A

PROOFS OF THEOREMS 4.1 AND 4.2

Theorems 4.1 and 4.2 both contain two claims – an almost sure convergence and an inequality. For better readability we will prove these claims in separate theorems. Theorem 4.1 follows directly from Theorems A.1 and A.3, and similarly Theorem 4.2 follows from Theorems A.2 and A.4. Additionally, Theorems A.1 and A.2 include the conditions stating when the inequalities become equalities. For Lemmas A.2 and A.4 and Theorems A.2 and A.4 we define $0/0 = 0$ and $c/0 = \infty$ for any $c > 0$.

Proofs of the inequalities

Lemma A.1. *Let X and Y be binary random variables and W a real-valued random variable, where X and W are independent conditional to Y. If $0<P^{\mathsf{X}}(1)<1$, then the following equality holds for any $x \in \{0,1\}$ and $w \in \mathbb{R}$:*

$$F^{\mathsf{W}|\mathsf{X}}(w|x) = F^{\mathsf{W}|\mathsf{Y}}(w|1) \cdot P^{\mathsf{Y}|\mathsf{X}}(1|x) + F^{\mathsf{W}|\mathsf{Y}}(w|0) \cdot \left(1 - P^{\mathsf{Y}|\mathsf{X}}(1|x)\right)$$

Proof. Due to the independence of X and W conditional to Y, we have for any $x, y \in \{0,1\}$ and $w \in \mathbb{R}$

$$\mathbf{P}(W{\leq}w|Y{=}y,\, X{=}x) = \mathbf{P}(W{\leq}w|Y{=}y)$$

Therefore,

$$\begin{aligned}
&F^{\mathsf{W}|\mathsf{X}}(w|x) = \mathbf{P}(W{\leq}w|X{=}x) = \\
&= \mathbf{P}(W{\leq}w,\, Y{=}1|X{=}x) + \mathbf{P}(W{\leq}w,\, Y{=}0|X{=}x) = \\
&= \mathbf{P}(W{\leq}w|Y{=}1,\, X{=}x)\,\mathbf{P}(Y{=}1|X{=}x) + \mathbf{P}(W{\leq}w|Y{=}0,\, X{=}x)\,\mathbf{P}(Y{=}0|X{=}x) = \\
&= \mathbf{P}(W{\leq}w|Y{=}1)\,\mathbf{P}(Y{=}1|X{=}x) + \mathbf{P}(W{\leq}w|Y{=}0)\left(1 - \mathbf{P}(Y{=}1|X{=}x)\right) = \\
&= F^{\mathsf{W}|\mathsf{Y}}(w|1)\,P^{\mathsf{Y}|\mathsf{X}}(1|x) + F^{\mathsf{W}|\mathsf{Y}}(w|0)\left(1 - P^{\mathsf{Y}|\mathsf{X}}(1|x)\right)
\end{aligned}$$

□

Theorem A.1. *Let X and Y be binary random variables and W a real-valued random variable, where X and W are independent conditional to Y. If $0<P^{\mathsf{X}}(1)<1$, then the following inequality holds:*

$$\mathsf{KS}^*(W|X) \leq \mathsf{ARC}(Y|X)$$

where

$$\mathsf{KS}^*(W|X) = \sup_{w\in\mathbb{R}} \left| F^{\mathsf{W}|\mathsf{X}}(w|1) - F^{\mathsf{W}|\mathsf{X}}(w|0) \right|$$
$$\mathsf{ARC}(Y|X) = \left| P^{\mathsf{Y}|\mathsf{X}}(1|1) - P^{\mathsf{Y}|\mathsf{X}}(1|0) \right|$$

The inequality becomes an equality if and only if $\mathsf{KS}^*(W|Y) = 1$.

Proof. The theorem follows easily from Lemma A.1:

$$\begin{aligned}
\mathsf{KS}^*(W|X) &= \sup_{w\in\mathbb{R}} \left| F^{\mathsf{W}|\mathsf{X}}(w|1) - F^{\mathsf{W}|\mathsf{X}}(w|0) \right| = \\
&\overset{\text{Lemma A.1}}{=} \sup_{w\in\mathbb{R}} \Big| F^{\mathsf{W}|\mathsf{Y}}(w|1)\, P^{\mathsf{Y}|\mathsf{X}}(1|1) + F^{\mathsf{W}|\mathsf{Y}}(w|0) \left(1 - P^{\mathsf{Y}|\mathsf{X}}(1|1)\right) - \\
&\qquad - F^{\mathsf{W}|\mathsf{Y}}(w|1)\, P^{\mathsf{Y}|\mathsf{X}}(1|0) - F^{\mathsf{W}|\mathsf{X}}(w|0) \left(1 - P^{\mathsf{Y}|\mathsf{X}}(1|0)\right) \Big| = \\
&= \sup_{w\in\mathbb{R}} \left| \left(P^{\mathsf{Y}|\mathsf{X}}(1|1) - P^{\mathsf{Y}|\mathsf{X}}(1|0)\right) \cdot \left(F^{\mathsf{W}|\mathsf{Y}}(w|1) - F^{\mathsf{W}|\mathsf{Y}}(w|0)\right) \right| = \\
&= \left| P^{\mathsf{Y}|\mathsf{X}}(1|1) - P^{\mathsf{Y}|\mathsf{X}}(1|0) \right| \cdot \sup_{w\in\mathbb{R}} \left| F^{\mathsf{W}|\mathsf{Y}}(w|1) - F^{\mathsf{W}|\mathsf{Y}}(w|0) \right| = \\
&= \mathsf{ARC}(Y|X) \cdot \mathsf{KS}^*(W|Y) \leq \mathsf{ARC}(Y|X)
\end{aligned}$$

The inequality becomes an equality if and only if $\mathsf{KS}^*(W|Y) = 1$. □

Lemma A.2. *Let* p, q *be real numbers with* $0 < q < p < 1$ *and* f, g *be functions with* $f, g : \mathbb{R} \to [0, \infty)$. *Then*

$$\sup_{t\in\mathbb{R}} \frac{p \cdot f(t) + (1-p) \cdot g(t)}{q \cdot f(t) + (1-q) \cdot g(t)} \leq \frac{p}{q}$$

where the equality holds if and only if $\sup_{t\in\mathbb{R}} \frac{f(t)}{g(t)} = \infty$.

Proof. For any $t \in \mathbb{R}$ the following holds:

$$\frac{p \cdot f(t) + (1-p) \cdot g(t)}{q \cdot f(t) + (1-q) \cdot g(t)} \leq \frac{p}{q} \Leftrightarrow \frac{qp \cdot f(t) + q(1-p) \cdot g(t)}{pq \cdot f(t) + p(1-q) \cdot g(t)} = \frac{pq(f(t) - g(t)) + q \cdot g(t)}{pq(f(t) - g(t)) + p \cdot g(t)} \leq 1$$

The latter inequality holds because $q < p$, and thus we have proved the inequality stated by the lemma. If $g(t) = 0$ then the inequality becomes an equality if and only if $f(t) > 0$. If $g(t) > 0$ then

$$\frac{pq(f(t) - g(t)) + q \cdot g(t)}{pq(f(t) - g(t)) + p \cdot g(t)} = \frac{pq(\frac{f(t)}{g(t)} - 1) + q}{pq(\frac{f(t)}{g(t)} - 1) + p}$$

which converges to 1 if and only if $\frac{f(t)}{g(t)} \to \infty$. The two cases can be joined under the condition $\sup_{t\in\mathbb{R}} \frac{f(t)}{g(t)} = \infty$, proving the lemma. □

Theorem A.2. *Let* X *and* Y *be binary random variables and* W *a real-valued random variable, where* X *and* W *are independent conditional to* Y. *If* $0 < P^{\mathsf{X}}(1) < 1$ *and*

$$0 < P^{\mathsf{Y}|\mathsf{X}}(1|0) < P^{\mathsf{Y}|\mathsf{X}}(1|1) < 1$$

then the following inequality holds:

$$\mathsf{Q}^*(W|X) \leq \mathsf{RR}(Y|X)$$

where

$$\mathsf{Q}^*(W|X) = \sup_{w\in\mathbb{R}} \frac{F^{\mathsf{W}|\mathsf{X}}(w|1)}{F^{\mathsf{W}|\mathsf{X}}(w|0)} \qquad \mathsf{RR}(Y|X) = \frac{P^{\mathsf{Y}|\mathsf{X}}(1|1)}{P^{\mathsf{Y}|\mathsf{X}}(1|0)}$$

The inequality becomes an equality if and only if $\mathsf{Q}^*(W|Y) = \infty$.

Proof. Applying Lemma A.1 and Lemma A.2 with

$$p = P^{\mathsf{Y}|\mathsf{X}}(1|1) \qquad q = P^{\mathsf{Y}|\mathsf{X}}(1|0) \qquad f(w) = F^{\mathsf{W}|\mathsf{Y}}(w|1) \qquad g(w) = F^{\mathsf{W}|\mathsf{Y}}(w|0)$$

we obtain the required inequality:

$$\begin{aligned}\mathsf{Q}^*(W|X) = \sup_{w\in\mathbb{R}} \frac{F^{\mathsf{W}|\mathsf{X}}(w|1)}{F^{\mathsf{W}|\mathsf{X}}(w|0)} = \\ \overset{\text{Lemma A.1}}{=} \sup_{w\in\mathbb{R}} \frac{F^{\mathsf{W}|\mathsf{Y}}(w|1)\cdot P^{\mathsf{Y}|\mathsf{X}}(1|1) + F^{\mathsf{W}|\mathsf{Y}}(w|0)\cdot(1-P^{\mathsf{Y}|\mathsf{X}}(1|1))}{F^{\mathsf{W}|\mathsf{Y}}(w|1)\cdot P^{\mathsf{Y}|\mathsf{X}}(1|0) + F^{\mathsf{W}|\mathsf{Y}}(w|0)\cdot(1-P^{\mathsf{Y}|\mathsf{X}}(1|0))} \leq \\ \overset{\text{Lemma A.2}}{\leq} \frac{P^{\mathsf{Y}|\mathsf{X}}(1|1)}{P^{\mathsf{Y}|\mathsf{X}}(1|0)} = \mathsf{RR}(Y|X)\end{aligned}$$

According to Lemma A.2 the last inequality becomes an equality if and only if

$$\sup_{w\in\mathbb{R}} \frac{F^{\mathsf{W}|\mathsf{Y}}(w|1)}{F^{\mathsf{W}|\mathsf{Y}}(w|0)} = \infty$$

which is by the notations the same as $\mathsf{Q}^*(W|Y) = \infty$. □

Proofs of the almost sure convergences

To prove almost sure convergences we will need Lemma A.3, which is proved using the strong law of large numbers, the classical Borel-Cantelli lemma [8] and the Glivenko-Cantelli theorem as stated by Devroye *et al.* [19]. Note that this version of the Glivenko-Cantelli theorem includes a concentration inequality which is required for proving Corollary A.3.

Corollary A.1 (of the strong law of large numbers [8])**.** *Let $X_1, X_2, \ldots$ be i.i.d. random variables distributed identically to a binary random variable X. Then for each $x \in \{0,1\}$*

$$\lim_{n\to\infty} \left| P_n^{\mathsf{X}}(x) - P^{\mathsf{X}}(x) \right| = 0 \quad \text{a.s.}$$

Corollary A.2 (of the first Borel-Cantelli lemma [8])**.** *Let $R_1, R_2, \ldots$ be real-valued random variables. If $\sum_{n=1}^{\infty} \mathbf{P}(R_n > \epsilon)$ converges for each $\epsilon > 0$, then $\lim_{n\to\infty} R_n = 0$ almost surely.*

Theorem (Glivenko-Cantelli [19])**.** *Let $W_1, W_2, \ldots, W_n$ be i.i.d. real-valued random variables with the cumulative distribution function $F^{\mathsf{W}}(w)$ and the empirical cumulative distribution function $F_n^{\mathsf{W}}(w)$. Then for each $\epsilon > 0$*

$$\mathbf{P}\left(\sup_{w\in\mathbb{R}} \left| F_n^{\mathsf{W}}(w) - F^{\mathsf{W}}(w) \right| > \epsilon \right) \leq 8(n+1)\exp\left[-\frac{n\epsilon^2}{32} \right]$$

In particular, by the first Borel-Cantelli lemma

$$\lim_{n\to\infty} \sup_{w\in\mathbb{R}} \left| F_n^{\mathsf{W}}(w) - F^{\mathsf{W}}(w) \right| = 0, \quad \text{a.s.}$$

Corollary A.3 (of the Glivenko-Cantelli theorem)**.** *Let $W_1, W_2, \ldots, W_n$ be i.i.d. real-valued random variables with cumulative distribution function $F^{\mathsf{W}}(w)$ and empirical cumulative distribution function $F_n^{\mathsf{W}}(w)$. Then for any $\alpha < 1/2$ the following holds:*

$$\lim_{n\to\infty} n^{\alpha} \sup_{w\in\mathbb{R}} \left| F_n^{\mathsf{W}}(w) - F^{\mathsf{W}}(w) \right| = 0, \quad \text{a.s.}$$

Proof. According to Corollary A.2 it is sufficient to prove the convergence of the series

$$\sum_{n=1}^{\infty} \mathbf{P}\left(n^{\alpha} \sup_{w\in\mathbb{R}} \left| F_n^{\mathsf{W}}(w) - F^{\mathsf{W}}(w) \right| > \epsilon \right)$$

for any $\epsilon > 0$. By the Glivenko-Cantelli theorem we obtain

$$\begin{aligned} \mathbf{P}\left(n^{\alpha} \sup_{w\in\mathbb{R}} \left| F_n^{\mathsf{W}}(w) - F^{\mathsf{W}}(w) \right| > \epsilon \right) &\leq 8(n+1) \exp\left[-\frac{n(\epsilon/n^{\alpha})^2}{32} \right] = \\ &= 8(n+1) \exp\left[-\frac{n^{1-2\alpha}\epsilon^2}{32} \right] \end{aligned}$$

The required series converges because $1 - 2\alpha > 0$. □

Lemma A.3. *Let $(X_i, W_i)_{i=1}^n$ be an i.i.d. sample from the joint distribution of a binary random variable X and a real-valued random variable W. If $0 < P^{\mathsf{X}}(1) < 1$, then for any $\alpha < 1/2$ and any $x \in \{0, 1\}$ the following holds:*

$$\lim_{n\to\infty} n^{\alpha} \sup_{w\in\mathbb{R}} \left| F_n^{\mathsf{W}|\mathsf{X}}(w|x) - F^{\mathsf{W}|\mathsf{X}}(w|x) \right| = 0, \quad \text{a.s.}$$

Proof. First note that it is enough to prove the almost sure convergence under the assumption

$$\lim_{n\to\infty} P_n^{\mathsf{X}}(x) = P^{\mathsf{X}}(x)$$

because this is an almost sure event according to Corollary A.1. As $N_n^{\mathsf{X}}(x) = n \cdot P_n^{\mathsf{X}}(x)$ and $P^{\mathsf{X}}(x) > 0$ then $N_n^{\mathsf{X}}(x)$ grows to infinity with $n \to \infty$.

Let $W_1', W_2', \ldots$ be the subsequence of $W_1, W_2, \ldots$ where we include all W_i for which $X_i = x$. For any n, the number of included values W_i among $W_1, \ldots, W_n$ is equal to $N_n^{\mathsf{X}}(x)$. Therefore, the following sets are equal

$$\left\{ W_i' \,\middle|\, 1 \leq i \leq N_n^{\mathsf{X}}(x) \right\} = \left\{ W_i \,\middle|\, 1 \leq i \leq n, X_i = x \right\}$$

and the proportion of values below any threshold $w \in \mathbb{R}$ must be the same in these two sets,

$$F_{N_n^{\mathsf{X}}(x)}^{\mathsf{W}'}(w) = F_n^{\mathsf{W}|\mathsf{X}}(w|x)$$

As the random variables $W_1, W_2, \ldots$ were i.i.d., then so must be the variables $W_1', W_2', \ldots$. Each of the variables W_i' has the cumulative distribution function

$$F^{\mathsf{W}'}(w) = F^{\mathsf{W}|\mathsf{X}}(w|x)$$

due to the condition $X_i = x$. The sequence $W_1', W_2', \ldots$ is infinite because $\lim_{n\to\infty} N_n^{\mathsf{X}}(x) = \infty$. Now we can write

$$\lim_{n\to\infty} n^{\alpha} \sup_{w\in\mathbb{R}} \left| F_n^{\mathsf{W}|\mathsf{X}}(w|x) - F^{\mathsf{W}|\mathsf{X}}(w|x) \right| = \frac{\lim_{n\to\infty} \left(N_n^{\mathsf{X}}(x) \right)^{\alpha} \sup_{w\in\mathbb{R}} \left| F_{N_n^{\mathsf{X}}(x)}^{\mathsf{W}'}(w) - F^{\mathsf{W}'}(w) \right|}{\lim_{n\to\infty} \left(\frac{N_n^{\mathsf{X}}(x)}{n} \right)^{\alpha}}$$

According to Corollary A.3 the numerator is almost surely zero and by our assumption $N_n^{\mathsf{X}}(x)/n = P_n^{\mathsf{X}}(x)$ converges to $P^{\mathsf{X}}(x) \neq 0$, proving the lemma. □

In the final two theorems we study only the association measures between W and X, and therefore we drop $(W|X)$ in the notation of measures, *e.g.* we write KS_n instead of $\mathsf{KS}_n(W|X)$. In addition, we introduce the following notations for $x=0$ and $x=1$:

$$\begin{aligned} F_n^x(w) &= F_n^{\mathsf{W}|\mathsf{X}}(w|x) \\ F^x(w) &= F^{\mathsf{W}|\mathsf{X}}(w|x) \\ \Delta_n^x &= \sup_{w\in\mathbb{R}} |F_n^x(w) - F^x(w)| \end{aligned}$$

Theorem A.3. *Let $(X_i, W_i)_{i=1}^n$ be an i.i.d. sample from the joint distribution of a binary random variable X and a real-valued random variable W. If $0 < P^{\mathsf{X}}(1) < 1$, then the following convergence occurs almost surely:*

$$\lim_{n\to\infty} \mathsf{KS}_n = \mathsf{KS}^* \quad \text{a.s.}$$

where

$$\begin{aligned} \mathsf{KS}_n &= \sup_{w\in\mathbb{R}} \left|F_n^1(w) - F_n^0(w)\right| \\ \mathsf{KS}^* &= \sup_{w\in\mathbb{R}} \left|F^1(w) - F^0(w)\right| \end{aligned}$$

Proof. Since $\Big|\,|a-b| - |c-d|\,\Big| \le |a-c| + |b-d|$ for any $a, b, c, d \in \mathbb{R}$, then for any $w \in \mathbb{R}$

$$\begin{aligned} &\Big|\left|F_n^1(w) - F_n^0(w)\right| - \left|F^1(w) - F^0(w)\right|\Big| \le \\ &\le \left|F_n^1(w) - F^1(w)\right| + \left|F_n^0(w) - F^0(w)\right| \end{aligned}$$

Therefore,

$$\begin{aligned} |\mathsf{KS}_n - \mathsf{KS}^*| &= \Big|\sup_{w\in\mathbb{R}}\left|F_n^1(w) - F_n^0(w)\right| - \sup_{w\in\mathbb{R}}\left|F^1(w) - F^0(w)\right|\Big| \le \\ &\le \sup_{w\in\mathbb{R}}\Big|\left|F_n^1(w) - F_n^0(w)\right| - \left|F^1(w) - F^0(w)\right|\Big| \le \\ &\le \sup_{w\in\mathbb{R}}\Big(\left|F_n^1(w) - F^1(w)\right| + \left|F_n^0(w) - F^0(w)\right|\Big) \le \\ &\le \sup_{w\in\mathbb{R}}\left|F_n^1(w) - F^1(w)\right| + \sup_{w\in\mathbb{R}}\left|F_n^0(w) - F^0(w)\right| \to 0 \quad \text{a.s.} \end{aligned}$$

where the convergence is due to Lemma A.3 with $\alpha = 0$. □

Lemma A.4. *Let $a, c, d \ge 0$ and $b > 0$. If $c/d < \infty$, then*

$$\left|\frac{a}{b} - \frac{c}{d}\right| \le \frac{|a-c|}{b} + \frac{c\cdot|b-d|}{b\cdot(b-|b-d|)}$$

Proof. If $d=0$, then also $c=0$, since $c/d < \infty$. In this case the inequality holds as both sides are equal to a/b. If $d>0$, then

$$\left|\frac{a}{b} - \frac{c}{d}\right| \le \left|\frac{a}{b} - \frac{c}{b}\right| + \left|\frac{c}{b} - \frac{c}{d}\right| = \frac{|a-c|}{b} + \frac{c\cdot|b-d|}{b\cdot d}$$

Now it remains to prove that $d \ge b - |b-d|$, which follows immediately from $b - d \le |b-d|$. □

Theorem A.4. *Let $(X_i, W_i)_{i=1}^n$ be an i.i.d. sample from the joint distribution of a binary random variable X and a real-valued random variable W. If $0 < P^{\mathsf{X}}(1) < 1$ and $\mathsf{Q}^* < \infty$, then for any β with $0 < \beta < \frac{1}{4}$ the following convergence occurs almost surely:*

$$\lim_{n\to\infty} \mathsf{Q}_n = \mathsf{Q}^* \quad \text{a.s.}$$

where

$$\mathsf{Q}_n = \sup_{w\in\mathcal{B}_n} \frac{F_n^1(w)}{F_n^0(w)} \qquad\qquad \mathsf{Q}^* = \sup_{w\in\mathbb{R}} \frac{F^1(w)}{F^0(w)}$$
$$\mathcal{B}_n = \left\{ w \in \mathbb{R} \,\middle|\, F_n^0(w) > n^{-\beta} \right\}$$

Proof. Let us first introduce some more notation:

$$\mathsf{Q}_n^* = \sup_{w\in\mathcal{B}_n} \frac{F^1(w)}{F^0(w)}$$

Since $|\mathsf{Q}_n - \mathsf{Q}^*| \le |\mathsf{Q}_n - \mathsf{Q}_n^*| + |\mathsf{Q}_n^* - \mathsf{Q}^*|$ then we can prove the theorem in two parts:

(A) $\quad |\mathsf{Q}_n - \mathsf{Q}_n^*| \overset{\text{a.s.}}{\to} 0$

(B) $\quad \mathsf{Q}_n^* \overset{\text{a.s.}}{\to} \mathsf{Q}^*$

(A) First we note that

$$|\mathsf{Q}_n - \mathsf{Q}_n^*| = \left| \sup_{w\in\mathcal{B}_n} \frac{F_n^1(w)}{F_n^0(w)} - \sup_{w\in\mathcal{B}_n} \frac{F^1(w)}{F^0(w)} \right| \le \sup_{w\in\mathcal{B}_n} \left| \frac{F_n^1(w)}{F_n^0(w)} - \frac{F^1(w)}{F^0(w)} \right|$$

As for any $w \in \mathcal{B}_n$ we have $F_n^0(w) > n^{-\beta} > 0$ and $F^1(w)/F^0(w) \le \mathsf{Q}^* < \infty$, then we can apply Lemma A.4 with $a = F_n^1(w)$, $b = F_n^0(w)$, $c = F^1(w)$, $d = F^0(w)$ and get

$$\left| \frac{F_n^1(w)}{F_n^0(w)} - \frac{F^1(w)}{F^0(w)} \right| \le \frac{\left|F_n^1(w) - F^1(w)\right|}{F_n^0(w)} + \frac{F^1(w)\left|F_n^0(w) - F^0(w)\right|}{F_n^0(w)\cdot\left(F_n^0(w) - |F_n^0(w) - F^0(w)|\right)} \le$$
$$\le \frac{\Delta_n^1}{n^{-\beta}} + \frac{1\cdot\Delta_n^0}{n^{-\beta}\cdot(n^{-\beta} - \Delta_n^0)} = n^\beta\Delta_n^1 + \frac{n^{2\beta}\Delta_n^0}{1 - n^\beta\Delta_n^0}$$

The required result follows now from Lemma A.3 with $\alpha = \beta$ and $\alpha = 2\beta$ since $\beta < 2\beta < \frac{1}{2}$:

$$|\mathsf{Q}_n - \mathsf{Q}_n^*| \le \sup_{w\in\mathcal{B}_n} \left| \frac{F_n^1(w)}{F_n^0(w)} - \frac{F^1(w)}{F^0(w)} \right| \le n^\beta\Delta_n^1 + \frac{n^{2\beta}\Delta_n^0}{1 - n^\beta\Delta_n^0} \to 0 \quad \text{a.s.}$$

(B) First we note that $\mathsf{Q}_n^* \le \mathsf{Q}^*$ since the supremum is taken of the same expression but in the case of Q^* over a larger set. Therefore,

$$\limsup_{n\to\infty} \mathsf{Q}_n^* \le \mathsf{Q}^*$$

It remains to show that $\liminf_{n\to\infty} \mathsf{Q}_n^* \ge \mathsf{Q}^*$ almost surely. Since $\mathsf{Q}^* < \infty$, it is possible to choose a sequence (w_m) such that $F^0(w_m) > 0$ for each m and

$$\mathsf{Q}^* = \sup_{w\in\mathbb{R}} \frac{F^1(w)}{F^0(w)} = \lim_{m\to\infty} \frac{F^1(w_m)}{F^0(w_m)}$$

Now according to Lemma A.3 with $\alpha = 0$ we get for each m

$$0 \leq \limsup_{n\to\infty} \left| \left(F_n^0(w_m) - n^{-\beta} \right) - F^0(w_m) \right| \leq$$
$$\leq \limsup_{n\to\infty} \left| F_n^0(w_m) - F^0(w_m) \right| + \limsup_{n\to\infty} n^{-\beta} \leq \limsup_{n\to\infty} \Delta_n^0 + \limsup_{n\to\infty} n^{-\beta} \overset{\text{a.s.}}{=} 0$$

and therefore,

$$\lim_{n\to\infty} \left(F_n^0(w_m) - n^{-\beta} \right) \overset{\text{a.s.}}{=} F^0(w_m)$$

Since $w_m \in \mathcal{B}_n$ if and only if $\operatorname{sgn}\left(F_n^0(w_m) - n^{-\beta} \right) = 1$ and since $F^0(w_m) > 0$, we get

$$\lim_{n\to\infty} I_{\mathcal{B}_n}(w_m) = \lim_{n\to\infty} \operatorname{sgn}\left(F_n^0(w_m) - n^{-\beta} \right) \overset{\text{a.s.}}{=} \operatorname{sgn} F^0(w_m) = 1$$

where $I_{\mathcal{B}_n}(w_m)$ is the indicator function with value 1 if $w_m \in \mathcal{B}_n$ and 0 otherwise. Now

$$\liminf_{n\to\infty} \mathsf{Q}_n^* = \liminf_{n\to\infty} \sup_{w\in\mathcal{B}_n} \frac{F^1(w)}{F^0(w)} = \liminf_{n\to\infty} \sup_{w\in\mathbb{R}} I_{\mathcal{B}_n}(w) \frac{F^1(w)}{F^0(w)} \geq$$
$$\geq \liminf_{n\to\infty} I_{\mathcal{B}_n}(w_m) \frac{F^1(w_m)}{F^0(w_m)} \overset{\text{a.s.}}{=} \frac{F^1(w_m)}{F^0(w_m)}$$

As this holds for every m, we finally get

$$\liminf_{n\to\infty} \mathsf{Q}_n^* \overset{\text{a.s.}}{\geq} \lim_{m\to\infty} \frac{F^1(w_m)}{F^0(w_m)} = \mathsf{Q}^*$$

which concludes the proof. □

Bibliography

[1] Aerts, S., Thijs, G., Coessens, B., Staes, M., Moreau, Y., Moor, B.D.: Toucan: deciphering the cis-regulatory logic of coregulated genes. Nucleic Acids Res. 31(6), 1753–64 (Mar 2003)

[2] Al-Shahrour, F., Diaz-Uriarte, R., Dopazo, J.: Discovering molecular functions significantly related to phenotypes by combining gene expression data and biological information. Bioinformatics 21(13), 2988–2993 (Apr 2005)

[3] Ashburner, M., Ball, C.A., Blake, J.A., Botstein, D., Butler, H., Cherry, J.M., Davis, A.P., Dolinski, K., Dwight, S.S., Eppig, J.T., Harris, M.A., Hill, D.P., Issel-Tarver, L., Kasarskis, A., Lewis, S., Matese, J.C., Richardson, J.E., Ringwald, M., Rubin, G.M., Sherlock, G.: Gene Ontology: tool for the unification of biology. Nat. Genet. 25(1), 25–9 (May 2000)

[4] Backes, C., Meese, E., Lenhof, H.P., Keller, A.: A dictionary on microRNAs and their putative target pathways. Nucleic Acids Res. 38(13), 4476–4486 (Jul 2010)

[5] Backes, C., Keller, A., Kuentzer, J., Kneissl, B., Comtesse, N., Elnakady, Y.A., Müller, R., Meese, E., Lenhof, H.P.: GeneTrail–advanced gene set enrichment analysis. Nucleic Acids Res. 35(Web Server issue), W186–92 (Jul 2007)

[6] Bailey, T.L., Elkan, C.: Fitting a mixture model by expectation maximization to discover motifs in biopolymers. Proceedings of the International Conference on Intelligent Systems for Molecular Biology 2, 28–36 (Jan 1994)

[7] Ben-Shaul, Y., Bergman, H., Soreq, H.: Identifying subtle interrelated changes in functional gene categories using continuous measures of gene expression. Bioinformatics 21(7), 1129–1137 (Nov 2004)

[8] Billingsley, P.: Probability and measure. Wiley, third edn. (1995)

[9] Billon, N., Kolde, R., Reimand, J., Monteiro, M.C., Kull, M., Peterson, H., Tretyakov, K., Adler, P., Wdziekonski, B., Vilo, J., Dani, C.: Comprehensive transcriptome analysis of mouse embryonic stem cell adipogenesis unravels new processes of adipocyte development. Genome Biol 11(8), R80 (Aug 2010)

[10] Bird, A.: Perceptions of epigenetics. Nature 447(7143), 396–398 (May 2007)

[11] Brazma, A., Jonassen, I., Vilo, J., Ukkonen, E.: Predicting gene regulatory elements in silico on a genomic scale. Genome Res. 8(11), 1202–15 (Dec 1998)

[12] Bryne, J.C., Valen, E., Tang, M.H.E., Marstrand, T., Winther, O., da Piedade, I., Krogh, A., Lenhard, B., Sandelin, A.: JASPAR, the open access database of transcription factor-binding profiles: new content and tools in the 2008 update. Nucleic Acids Res. 36(Database issue), D102–6 (Jan 2008)

[13] Bucher, P., Bryan, B.: Signal search analysis: a new method to localize and characterize functionally important DNA sequences. Nucleic Acids Res. 12(1 Pt 1), 287–305 (Jan 1984)

[14] Cardon, L.R., Burge, C., Clayton, D.A., Karlin, S.: Pervasive CpG suppression in animal mitochondrial genomes. Proc Natl Acad Sci USA 91(9), 3799–803 (Apr 1994)

[15] Casella, G., Berger, R.L.: Statistical Inference. Duxbury Press, second edn. (2002)

[16] Chang, L.W., Nagarajan, R., Magee, J.A., Milbrandt, J., Stormo, G.D.: A systematic model to predict transcriptional regulatory mechanisms based on overrepresentation of transcription factor binding profiles. Genome Res. 16(3), 405–13 (Mar 2006)

[17] Das, M.K., Dai, H.K.: A survey of DNA motif finding algorithms. BMC Bioinformatics 8 Suppl 7, S21 (Jan 2007)

[18] Day, W., Edelsbrunner, H.: Efficient algorithms for agglomerative hierarchical clustering methods. Journal of classification 1(1), 7–24 (1984)

[19] Devroye, L., Györfi, L., Lugosi, G.: A probabilistic theory of pattern recognition. Springer Verlag (1996)

[20] Doss, M.X., Wagh, V., Schulz, H., Kull, M., Kolde, R., Pfannkuche, K., Nolden, T., Himmelbauer, H., Vilo, J., Hescheler, J., Sachinidis, A.: Global

transcriptomic analysis of murine embryonic stem cell-derived brachyury$^+$ (T) cells. Genes to Cells 15(3), 209–228 (Feb 2010)

[21] Draghici, S., Khatri, P., Martins, R.P., Ostermeier, G.C., Krawetz, S.A.: Global functional profiling of gene expression. Genomics 81(2), 98–104 (Feb 2003)

[22] Eisen, M.B., Spellman, P.T., Brown, P.O., Botstein, D.: Cluster analysis and display of genome-wide expression patterns. Proc Natl Acad Sci USA 95(25), 14863–8 (Dec 1998)

[23] Ernst, J., Kheradpour, P., Mikkelsen, T.S., Shoresh, N., Ward, L.D., Epstein, C.B., Zhang, X., Wang, L., Issner, R., Coyne, M., Ku, M., Durham, T., Kellis, M., Bernstein, B.E.: Mapping and analysis of chromatin state dynamics in nine human cell types. Nature 473(7345), 43–9 (May 2011)

[24] Fernández-Ramires, R., Solé, X., Cecco, L.D., Llort, G., Cazorla, A., Bonifaci, N., Garcia, M.J., Caldés, T., Blanco, I., Gariboldi, M., Pierotti, M.A., Pujana, M.A., Benítez, J., Osorio, A.: Gene expression profiling integrated into network modelling reveals heterogeneity in the mechanisms of BRCA1 tumorigenesis. Br J Cancer 101(8), 1469–80 (Oct 2009)

[25] Field, Y., Sharon, E., Segal, E.: How transcription factors identify regulatory sites in genomic sequence. In: Harris, J.R., Hughes, T.R. (eds.) A Handbook of Transcription Factors, Subcellular Biochemistry, vol. 52. Springer (2011)

[26] Frith, M.C., Fu, Y., Yu, L., Chen, J.F., Hansen, U., Weng, Z.: Detection of functional DNA motifs via statistical over-representation. Nucleic Acids Res. 32(4), 1372–81 (Jan 2004)

[27] Guruceaga, E., Segura, V., Corrales, F.J., Rubio, A.: FactorY, a bioinformatic resource for genome-wide promoter analysis. Comput Biol Med 39(4), 385–7 (Apr 2009)

[28] Hamosh, A., Scott, A.F., Amberger, J.S., Bocchini, C.A., McKusick, V.A.: Online Mendelian Inheritance in Man (OMIM), a knowledgebase of human genes and genetic disorders. Nucleic Acids Res. 33(Database issue), D514–D517 (Dec 2005)

[29] van Helden, J., André, B., Collado-Vides, J.: Extracting regulatory sites from the upstream region of yeast genes by computational analysis of oligonucleotide frequencies. J. Mol. Biol. 281(5), 827–42 (Aug 1998)

[30] Hestand, M.S., van Galen, M., Villerius, M.P., van Ommen, G.J.B., den Dunnen, J.T., 't Hoen, P.A.C.: CORE_TF: a user-friendly interface to identify evolutionary conserved transcription factor binding sites in sets of co-regulated genes. BMC Bioinformatics 9, 495 (Jan 2008)

[31] Ho Sui, S.J., Mortimer, J.R., Arenillas, D.J., Brumm, J., Walsh, C.J., Kennedy, B.P., Wasserman, W.W.: oPOSSUM: identification of over-represented transcription factor binding sites in co-expressed genes. Nucleic Acids Res. 33(10), 3154–64 (Jan 2005)

[32] Huang, D.W., Sherman, B.T., Lempicki, R.A.: Bioinformatics enrichment tools: paths toward the comprehensive functional analysis of large gene lists. Nucleic Acids Res 37(1), 1–13 (Jan 2009)

[33] Huynh, X., Guillet, F., Blanchard, J., Kuntz, P., Briand, H., Gras, R.: A graph-based clustering approach to evaluate interestingness measures: a tool and a comparative study. Quality Measures in Data Mining pp. 25–50 (2007)

[34] Jain, A., Murty, M., Flynn, P.: Data clustering: a review. ACM Computing Surveys (CSUR) 31(3), 264–323 (1999)

[35] Kanehisa, M., Goto, S., Furumichi, M., Tanabe, M., Hirakawa, M.: KEGG for representation and analysis of molecular networks involving diseases and drugs. Nucleic Acids Res. 38(Database issue), D355–60 (Jan 2010)

[36] Kerr, M.K., Martin, M., Churchill, G.A.: Analysis of variance for gene expression microarray data. J Comput Biol 7(6), 819–37 (Jan 2000)

[37] Khatri, P., Draghici, S.: Ontological analysis of gene expression data: current tools, limitations, and open problems. Bioinformatics 21(18), 3587–3595 (Aug 2005)

[38] Krushevskaya, D., Peterson, H., Reimand, J., Kull, M., Vilo, J.: VisHiC – hierarchical functional enrichment analysis of microarray data. Nucl. Acids Res. 37(Web Server issue), W587–92 (Jul 2009)

[39] Kull, M., Vilo, J.: Fast approximate hierarchical clustering using similarity heuristics. BioData Mining 1(1), 9 (Sep 2008)

[40] Kull, M., Tretyakov, K., Vilo, J.: An evolutionary model of DNA substring distribution. In: Elomaa, T., Mannila, H., Orponen, P. (eds.) Algorithms and Applications, Essays Dedicated to Esko Ukkonen on the Occasion of His 60th Birthday, Lecture Notes in Computer Science, vol. 6060, pp. 147–157. Springer (2010)

[41] Lilienfeld, A.M., Lilienfeld, D.E.: Foundations of Epidemiology. Oxford University Press, second edn. (1980)

[42] Margulies, M., Egholm, M., Altman, W.E., Attiya, S., Bader, J.S., Bemben, L.A., Berka, J., Braverman, M.S., Chen, Y.J., Chen, Z., Dewell, S.B., Du, L., Fierro, J.M., Gomes, X.V., Godwin, B.C., He, W., Helgesen, S., Ho, C.H., Ho, C.H., Irzyk, G.P., Jando, S.C., Alenquer, M.L.I., Jarvie, T.P., Jirage, K.B., Kim, J.B., Knight, J.R., Lanza, J.R., Leamon, J.H., Lefkowitz, S.M., Lei, M., Li, J., Lohman, K.L., Lu, H., Makhijani, V.B., McDade, K.E., McKenna, M.P., Myers, E.W., Nickerson, E., Nobile, J.R., Plant, R., Puc, B.P., Ronan, M.T., Roth, G.T., Sarkis, G.J., Simons, J.F., Simpson, J.W., Srinivasan, M., Tartaro, K.R., Tomasz, A., Vogt, K.A., Volkmer, G.A., Wang, S.H., Wang, Y., Weiner, M.P., Yu, P., Begley, R.F., Rothberg, J.M.: Genome sequencing in microfabricated high-density picolitre reactors. Nature 437(7057), 376–80 (Sep 2005)

[43] Marstrand, T.T., Frellsen, J., Moltke, I., Thiim, M., Valen, E., Retelska, D., Krogh, A.: Asap: a framework for over-representation statistics for transcription factor binding sites. PLoS ONE 3(2), e1623 (Jan 2008)

[44] Matys, V., Kel-Margoulis, O.V., Fricke, E., Liebich, I., Land, S., Barre-Dirrie, A., Reuter, I., Chekmenev, D., Krull, M., Hornischer, K., Voss, N., Stegmaier, P., Lewicki-Potapov, B., Saxel, H., Kel, A.E., Wingender, E.: TRANSFAC and its module TRANSCompel: transcriptional gene regulation in eukaryotes. Nucleic Acids Res. 34(Database issue), D108–10 (Jan 2006)

[45] McLeay, R.C., Bailey, T.L.: Motif Enrichment Analysis: a unified framework and an evaluation on ChIP data. BMC Bioinformatics 11, 165 (Jan 2010)

[46] Meng, G., Mosig, A., Vingron, M.: A computational evaluation of over-representation of regulatory motifs in the promoter regions of differentially expressed genes. BMC Bioinformatics 11, 267 (Jan 2010)

[47] Mewes, H.W., Albermann, K., Bähr, M., Frishman, D., Gleissner, A., Hani, J., Heumann, K., Kleine, K., Maierl, A., Oliver, S.G., Pfeiffer, F., Zollner, A.: Overview of the yeast genome. Nature 387(6632 Suppl), 7–65 (May 1997)

[48] Mortazavi, A., Williams, B.A., Mccue, K., Schaeffer, L., Wold, B.: Mapping and quantifying mammalian transcriptomes by RNA-Seq. Nat Methods 5(7), 621–628 (Jul 2008)

[49] Pesole, G., Prunella, N., Liuni, S., Attimonelli, M., Saccone, C.: WORDUP: an efficient algorithm for discovering statistically significant patterns in DNA sequences. Nucleic Acids Res. 20(11), 2871–5 (Jun 1992)

[50] Pique-Regi, R., Degner, J.F., Pai, A.A., Gaffney, D.J., Gilad, Y., Pritchard, J.K.: Accurate inference of transcription factor binding from DNA sequence and chromatin accessibility data. Genome Research 21(3), 447–455 (Mar 2011)

[51] Qin, Z.S., Yu, J., Shen, J., Maher, C.A., Hu, M., Kalyana-Sundaram, S., Yu, J., Chinnaiyan, A.M.: HPeak: an HMM-based algorithm for defining read-enriched regions in ChIP-Seq data. BMC Bioinformatics 11, 369 (Jan 2010)

[52] Quackenbush, J.: Computational analysis of microarray data. Nat Rev Genet 2(6), 418–27 (Jun 2001)

[53] Redhead, E., Bailey, T.: Discriminative motif discovery in DNA and protein sequences using the DEME algorithm. BMC Bioinformatics 8(1), 385 (Oct 2007)

[54] Reimand, J., Kull, M., Peterson, H., Hansen, J., Vilo, J.: g:Profiler–a web-based toolset for functional profiling of gene lists from large-scale experiments. Nucleic Acids Res. 35(Web Server issue), W193–200 (Jul 2007)

[55] Roider, H.G., Manke, T., O'Keeffe, S., Vingron, M., Haas, S.A.: PASTAA: identifying transcription factors associated with sets of co-regulated genes. Bioinformatics 25(4), 435–42 (Feb 2009)

[56] Schena, M., Shalon, D., Davis, R.W., Brown, P.O.: Quantitative monitoring of gene expression patterns with a complementary DNA microarray. Science 270(5235), 467–70 (Oct 1995)

[57] Shorack, G.R., Wellner, J.A.: Empirical Processes with Applications to Statistics. Wiley (1986)

[58] Smith, M.U., Adkison, L.R.: Updating the model definition of the gene in the modern genomic era with implications for instruction. Sci & Educ 19(1), 1–20 (Jan 2010)

[59] Subramanian, A., Tamayo, P., Mootha, V.K., Mukherjee, S., Ebert, B.L., Gillette, M.A., Paulovich, A., Pomeroy, S.L., Golub, T.R., Lander, E.S., Mesirov, J.P.: Gene set enrichment analysis: a knowledge-based approach for interpreting genome-wide expression profiles. Proc Natl Acad Sci USA 102(43), 15545–50 (Oct 2005)

[60] Tan, P., Kumar, V., Srivastava, J.: Selecting the right objective measure for association analysis. Information Systems 29(4), 293–313 (2004)

[61] Tavazoie, S., Hughes, J.D., Campbell, M.J., Cho, R.J., Church, G.M.: Systematic determination of genetic network architecture. Nat. Genet. 22(3), 281 (Jul 1999)

[62] Thijs, G., Lescot, M., Marchal, K., Rombauts, S., Moor, B.D., Rouzé, P., Moreau, Y.: A higher-order background model improves the detection of promoter regulatory elements by Gibbs sampling. Bioinformatics 17(12), 1113–22 (Dec 2001)

[63] Vingron, M., Brazma, A., Coulson, R., van Helden, J., Manke, T., Palin, K., Sand, O., Ukkonen, E.: Integrating sequence, evolution and functional genomics in regulatory genomics. Genome Biol 10(1), 202 (Jan 2009)

[64] Yang, X., Schadt, E.E., Wang, S., Wang, H., Arnold, A.P., Ingram-Drake, L., Drake, T.A., Lusis, A.J.: Tissue-specific expression and regulation of sexually dimorphic genes in mice. Genome Res. 16(8), 995–1004 (Aug 2006)

[65] Zambelli, F., Pesole, G., Pavesi, G.: Pscan: finding over-represented transcription factor binding site motifs in sequences from co-regulated or co-expressed genes. Nucleic Acids Res. 37(Web Server issue), W247–52 (Jul 2009)

[66] Zezula, P., Amato, G., Dohnal, V., Batko, M.: Similarity Search: the metric space approach, Advances in Database Systems, vol. 32. Springer (2006)

ACKNOWLEDGMENTS

First and above all I thank my dear Tiina and Esme Luise for their smile which changes the world every day. I am grateful to my parents and grandparents who got me interested in biology and mathematics, and to all my family for my positive thinking. My exploration of science has been guided by my supervisor prof. Jaak Vilo and my father prof. Kalevi Kull – to them belongs my deepest gratitude. I am grateful to all the co-authors I have had the honour to work with, to Konstantin Tretjakov and Swen Laur for their useful advice on improving the dissertation, and to Raivo Kolde, Tauno Metsalu and Märt Möls for guiding me to the appropriate statistical resources. My special thanks go to Jüri Lember for his advanced course on probability theory and his invaluable suggestions concerning the statistical parts of the dissertation. Last but not least, I thank my colleagues in the BIIT-group and friends in Estonian computer science for the good company, inspiration and support.

The work for this dissertation was financially partly supported by the Estonian Doctoral School of Information and Communication Technology (IKTDK), Centre of Excellence in Computer Science (EXCS), Estonian Target Funding grant no. SF0182712s06, Estonian Science Foundation grants no. 5722 (DMMA), 5724 (BiGeR), 7437 (MEM) and EU Framework 6 projects ATD (LSHG-CT-2003-503329), FunGenES (LSHG-CT-2003-503494), ENFIN (LSHG-CT-2005-518254) and COBRED (LSHB-CT-2007-037730). I also acknowledge the individual fellowships from the Marie Curie Biostar programme and the Tiger University programme of the Estonian Information Technology Foundation.

KOKKUVÕTE (SUMMARY IN ESTONIAN)

STATISTILINE RIKASTATUSE ANALÜÜS GEENIREGULATSIOONI UURIMISEKS LOODUD ALGORITMIDES

Üha suurema hulga organismide genoomide sekveneerimisega on meil tekkimas päris hea ülevaade geenide mitmekesisusest. Geenid, mis on põhiliseks ehitusinstruktsiooniks elule Maal, määravad ära selle, *kuidas* toota vajalikke RNA- ja valgumolekule, kuid mitte seda, *millal* ja *kui palju* toota. *Geeniregulatsioon* on mõiste, mis viitab suurele hulgale geeniproduktide tootmise hulka ja ajastust mõjutavatele mehhanismidele. Häired üheainsa geeni regulatsioonis võivad inimesel põhjustada mitmeid haigusi ja sündroome.

Käesoleva dissertatsiooni eesmärk on edasi arendada ja rakendada geeniregulatsiooni uurimiseks loodud algoritme, mis on seotud *statistilise rikastatusanalüüsiga*. Statistiline rikastatusanalüüs on hulk andmeanalüüsi meetodeid, mis uurivad, kas ja mil määral on andmed mingi suuruse poolest rikastatud. Seda analüüsi saab rakendada, kui meil on vaadeldava suuruse arvulise väärtuse suhtes mingi ootus, ja sellisel juhul viitab rikastatus juhtumile, kus tegelik väärtus osutub oodatust oluliselt suuremaks. Rikastatusanalüüsi on kasutatud laialdaselt bioinformaatikas, et uurida geenide ja muude bioloogiliste objektide vahelisi seoseid, kombineerides andmeid eksperimentidest ja bioloogilistest andmebaasidest. Dissertatsioon annab üldise formaalse definitsiooni rikastatuse kohta, mida meile teadaolevalt varem tehtud pole.

Geeniregulatsiooni uurimiseks kasutatakse sageli funktsionaalse rikastatuse analüüsi ning motiivide rikastatuse analüüsi. Esimene neist uurib, kas etteantud

geenide kogum sisaldab oodatavast rohkem ühe ja sama funktiooniga annoteeritud geene. Teine uurib aga seda, kas etteantud motiiv esineb oodatavast rohkemal määral mingis bioloogilises sekventsis. Dissertatsioon toob esile selle, et mõlemal juhul on sisuliselt tegemist bioloogiliste objektide omaduste vahelise assotsiatsiooniga, ning esitab formaalsele rikastatuse definitsioonile vastaval kujul kaks tuntud statistilist meetodit assotsiatsiooni tuvastamiseks – Fisheri täpse testi ning Kolmogorov-Smirnovi testi.

Järgnevas baseerub dissertatsioon viiele artiklile, mis on ära toodud lk. 67–145.

Esimeses artiklis („*Fast approximate hierarchical clustering using similarity heuristics*") on geenikespressiooni andmete hierarhilise klasterdamise jaoks välja arendatud kiire ligikaudne algoritm, mis kasutab lähedusheuristikuid. Kiirus on saavutatud kauguste arvutamisega vaid osade geenipaaride vahel, kusjuures klasterduse kvaliteeti tõstab heuristik, mis leiab lähedaste paaridega rikastatud alamhulga kõigist geenipaaridest.

Teises artiklis („*VisHiC–hierarchical functional enrichment analysis of microarray data*") on välja arendatud interaktiivne tarkvara geenikespressiooniandmete visualiseerimiseks. Tarkvara rakendab esimeses artiklis loodud algoritmist saadud hierarhilise klasterduse kõigile klastritele funktsionaalse rikastatuse analüüsi ning tõstab visuaalselt esile sarnase funktsiooni ja ekspressiooniga geenide grupid.

Kolmandas ja neljandas artiklis („*Global transcriptomic analysis of murine embryonic stem cell-derived brachyury$^+$ (T) cells*" ja „*Comprehensive transcriptome analysis of mouse embryonic stem cell adipogenesis unravels new processes of adipocyte development*") on uuritud hiire embrüonaalsete tüvirakkude diferentseerumist ning käesoleva dissertatsiooni autori panus seisnes promootorpiirkondade regulatoorse rikastatuse analüüsi kavandamises ja teostamises. Dissertatsioonis on tõestatud, et analüüsis kasutatud uudne binaarse ja reaalarvulise tunnuse vahelise assotsiatsiooni mõõt on ligikaudseks alumiseks tõkkeks reguleerituse suhtelisele riskile (*relative risk*). See tulemus lisab kindlust, et mõlemas artiklis on eksperimentaalselt saadud geenigruppide regulaatoriteks ennustatud transkriptsioonifaktorite seas vähe valepositiivseid.

Viiendas artiklis („*An evolutionary model of DNA substring distribution*") on loodud evolutsiooniline mudel, mis määrab lähtuvalt etteantud mutatsioonisagedustest ja sobivusfunktsioonist oodatava DNA alamstringide sagedusjaotuse. Mudelit saab potentsiaalselt rakendada taustana inkrementaalsel motiiviotsingul, kus iteratiivselt otsitakse uusi motiive ja täiendatakse taustamudelit juba leitud motiividega.

ORIGINAL PUBLICATIONS

CURRICULUM VITAE

Personal data

Name	Meelis Kull
Birth	February 21, 1980 Tartu, Estonia
Citizenship	Estonian
Languages	Estonian, English, Russian, German

Education

2004–	University of Tartu, Ph.D. candidate
2002–2004	University of Tartu, M.Sc. in Computer Science
1998–2002	University of Tartu, B.Sc. *cum laude* in Computer Science
1986–1998	Tartu Miina Härma Gymnasium, secondary education

Employment

2008–	University of Tartu, researcher
2005–2008	University of Tartu, extraordinary researcher
2004–2006	Estonian Biocentre, programmer
2006–2011	Quretec Ltd, researcher
2002–2006	EGeen Ltd, researcher

ELULOOKIRJELDUS

Isikuandmed

Nimi	Meelis Kull
Sünniaeg ja -koht	21. veebruar 1980 Tartu, Eesti
Kodakondsus	Eesti
Keelteoskus	eesti, inglise, vene, saksa

Haridustee

2004–	Tartu Ülikool, doktorant
2002–2004	Tartu Ülikool, MSc informaatikas
1998–2002	Tartu Ülikool, BSc *cum laude* informaatikas
1986–1998	Tartu Miina Härma Gümnaasium, keskharidus

Teenistuskäik

2008–	Tartu Ülikool, teadur
2005–2008	Tartu Ülikool, erakorraline teadur
2004–2006	Eesti Biokeskus, programmeerija
2006–2011	OÜ Quretec, teadur
2002–2006	AS EGeen, teadur

DISSERTATIONES MATHEMATICAE UNIVERSITATIS TARTUENSIS

1. **Mati Heinloo.** The design of nonhomogeneous spherical vessels, cylindrical tubes and circular discs. Tartu, 1991, 23 p.
2. **Boris Komrakov.** Primitive actions and the Sophus Lie problem. Tartu, 1991, 14 p.
3. **Jaak Heinloo.** Phenomenological (continuum) theory of turbulence. Tartu, 1992, 47 p.
4. **Ants Tauts.** Infinite formulae in intuitionistic logic of higher order. Tartu, 1992, 15 p.
5. **Tarmo Soomere.** Kinetic theory of Rossby waves. Tartu, 1992, 32 p.
6. **Jüri Majak.** Optimization of plastic axisymmetric plates and shells in the case of Von Mises yield condition. Tartu, 1992, 32 p.
7. **Ants Aasma.** Matrix transformations of summability and absolute summability fields of matrix methods. Tartu, 1993, 32 p.
8. **Helle Hein.** Optimization of plastic axisymmetric plates and shells with piece-wise constant thickness. Tartu, 1993, 28 p.
9. **Toomas Kiho.** Study of optimality of iterated Lavrentiev method and its generalizations. Tartu, 1994, 23 p.
10. **Arne Kokk.** Joint spectral theory and extension of non-trivial multiplicative linear functionals. Tartu, 1995, 165 p.
11. **Toomas Lepikult.** Automated calculation of dynamically loaded rigid-plastic structures. Tartu, 1995, 93 p, (in Russian).
12. **Sander Hannus.** Parametrical optimization of the plastic cylindrical shells by taking into account geometrical and physical nonlinearities. Tartu, 1995, 74 p, (in Russian).
13. **Sergei Tupailo.** Hilbert's epsilon-symbol in predicative subsystems of analysis. Tartu, 1996, 134 p.
14. **Enno Saks.** Analysis and optimization of elastic-plastic shafts in torsion. Tartu, 1996, 96 p.
15. **Valdis Laan.** Pullbacks and flatness properties of acts. Tartu, 1999, 90 p.
16. **Märt Põldvere.** Subspaces of Banach spaces having Phelps' uniqueness property. Tartu, 1999, 74 p.
17. **Jelena Ausekle.** Compactness of operators in Lorentz and Orlicz sequence spaces. Tartu, 1999, 72 p.
18. **Krista Fischer.** Structural mean models for analyzing the effect of compliance in clinical trials. Tartu, 1999, 124 p.

19. **Helger Lipmaa.** Secure and efficient time-stamping systems. Tartu, 1999, 56 p.
20. **Jüri Lember.** Consistency of empirical k-centres. Tartu, 1999, 148 p.
21. **Ella Puman.** Optimization of plastic conical shells. Tartu, 2000, 102 p.
22. **Kaili Müürisep.** Eesti keele arvutigrammatika: süntaks. Tartu, 2000, 107 lk.
23. **Varmo Vene.** Categorical programming with inductive and coinductive types. Tartu, 2000, 116 p.
24. **Olga Sokratova.** Ω-rings, their flat and projective acts with some applications. Tartu, 2000, 120 p.
25. **Maria Zeltser.** Investigation of double sequence spaces by soft and hard analitical methods. Tartu, 2001, 154 p.
26. **Ernst Tungel.** Optimization of plastic spherical shells. Tartu, 2001, 90 p.
27. **Tiina Puolakainen.** Eesti keele arvutigrammatika: morfoloogiline ühestamine. Tartu, 2001, 138 p.
28. **Rainis Haller.** *M(r,s)*-inequalities. Tartu, 2002, 78 p.
29. **Jan Villemson.** Size-efficient interval time stamps. Tartu, 2002, 82 p.
30. **Eno Tõnisson.** Solving of expession manipulation exercises in computer algebra systems. Tartu, 2002, 92 p.
31. **Mart Abel.** Structure of Gelfand-Mazur algebras. Tartu, 2003. 94 p.
32. **Vladimir Kuchmei.** Affine completeness of some ockham algebras. Tartu, 2003. 100 p.
33. **Olga Dunajeva.** Asymptotic matrix methods in statistical inference problems. Tartu 2003. 78 p.
34. **Mare Tarang.** Stability of the spline collocation method for volterra integro-differential equations. Tartu 2004. 90 p.
35. **Tatjana Nahtman.** Permutation invariance and reparameterizations in linear models. Tartu 2004. 91 p.
36. **Märt Möls.** Linear mixed models with equivalent predictors. Tartu 2004. 70 p.
37. **Kristiina Hakk.** Approximation methods for weakly singular integral equations with discontinuous coefficients. Tartu 2004, 137 p.
38. **Meelis Käärik.** Fitting sets to probability distributions. Tartu 2005, 90 p.
39. **Inga Parts.** Piecewise polynomial collocation methods for solving weakly singular integro-differential equations. Tartu 2005, 140 p.
40. **Natalia Saealle.** Convergence and summability with speed of functional series. Tartu 2005, 91 p.
41. **Tanel Kaart.** The reliability of linear mixed models in genetic studies. Tartu 2006, 124 p.
42. **Kadre Torn.** Shear and bending response of inelastic structures to dynamic load. Tartu 2006, 142 p.

43. **Kristel Mikkor.** Uniform factorisation for compact subsets of Banach spaces of operators. Tartu 2006, 72 p.
44. **Darja Saveljeva.** Quadratic and cubic spline collocation for Volterra integral equations. Tartu 2006, 117 p.
45. **Kristo Heero.** Path planning and learning strategies for mobile robots in dynamic partially unknown environments. Tartu 2006, 123 p.
46. **Annely Mürk.** Optimization of inelastic plates with cracks. Tartu 2006. 137 p.
47. **Annemai Raidjõe.** Sequence spaces defined by modulus functions and superposition operators. Tartu 2006, 97 p.
48. **Olga Panova.** Real Gelfand-Mazur algebras. Tartu 2006, 82 p.
49. **Härmel Nestra.** Iteratively defined transfinite trace semantics and program slicing with respect to them. Tartu 2006, 116 p.
50. **Margus Pihlak.** Approximation of multivariate distribution functions. Tartu 2007, 82 p.
51. **Ene Käärik.** Handling dropouts in repeated measurements using copulas. Tartu 2007, 99 p.
52. **Artur Sepp.** Affine models in mathematical finance: an analytical approach. Tartu 2007, 147 p.
53. **Marina Issakova.** Solving of linear equations, linear inequalities and systems of linear equations in interactive learning environment. Tartu 2007, 170 p.
54. **Kaja Sõstra.** Restriction estimator for domains. Tartu 2007, 104 p.
55. **Kaarel Kaljurand.** Attempto controlled English as a Semantic Web language. Tartu 2007, 162 p.
56. **Mart Anton.** Mechanical modeling of IPMC actuators at large deformations. Tartu 2008, 123 p.
57. **Evely Leetma.** Solution of smoothing problems with obstacles. Tartu 2009, 81 p.
58. **Ants Kaasik.** Estimating ruin probabilities in the Cramér-Lundberg model with heavy-tailed claims. Tartu 2009, 139 p.
59. **Reimo Palm.** Numerical Comparison of Regularization Algorithms for Solving Ill-Posed Problems. Tartu 2010, 105 p.
60. **Indrek Zolk.** The commuting bounded approximation property of Banach spaces. Tartu 2010, 107 p.
61. **Jüri Reimand.** Functional analysis of gene lists, networks and regulatory systems. Tartu 2010, 153 p.
62. **Ahti Peder.** Superpositional Graphs and Finding the Description of Structure by Counting Method. Tartu 2010, 87 p.
63. **Marek Kolk.** Piecewise Polynomial Collocation for Volterra Integral Equations with Singularities. Tartu 2010, 134 p.

64. **Vesal Vojdani.** Static Data Race Analysis of Heap-Manipulating C Programs. Tartu 2010, 137 p.
65. **Larissa Roots.** Free vibrations of stepped cylindrical shells containing cracks. Tartu 2010, 94 p.
66. **Mark Fišel.** Optimizing Statistical Machine Translation via Input Modification. Tartu 2011, 104 p.
67. **Margus Niitsoo**. Black-box Oracle Separation Techniques with Applications in Time-stamping. Tartu 2011, 174 p.
68. **Olga Liivapuu.** Graded q-differential algebras and algebraic models in noncommutative geometry. Tartu 2011, 112 p.
69. **Aleksei Lissitsin.** Convex approximation properties of Banach spaces. Tartu 2011, 107 p.
70. **Lauri Tart.** Morita equivalence of partially ordered semigroups. Tartu 2011, 101 p.
71. **Siim Karus.** Maintainability of XML Transformations. Tartu 2011, 142 p.
72. **Margus Treumuth.** A Framework for Asynchronous Dialogue Systems: Concepts, Issues and Design Aspects. Tartu 2011, 95 p.
73. **Dmitri Lepp.** Solving simplification problems in the domain of exponents, monomials and polynomials in interactive learning environment T-algebra. Tartu 2011, 202 p.