

JÜRI REIMAND

Functional analysis of gene lists,
networks and regulatory systems

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

Dissertation is accepted for the commencement of the degree of Doctor of Philosophy (PhD) on May 10th, 2010 by the Council of the Institute of Computer Science, University of Tartu.

Supervisor:

Prof. PhD.	Jaak Vilo Institute of Computer Science University of Tartu Tartu, Estonia
------------	---

Opponents:

Prof. D.Sc. Tech.	Samuel Kaski Aalto University School of Science and Technology Department of Information and Computer Science Helsinki, Finland
-------------------	--

Prof. Cand. Biol.	Toomas Neuman Tallinn University of Technology Centre for Biology of Integrated Systems Tallinn, Estonia
-------------------	---

The public defense will take place on June 17th, 2010 at 16:15 at the Institute of Computer Science, University of Tartu, at Liivi 2-404.

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

ISSN 1024-4212
ISBN 978-9949-19-368-4 (trükis)
ISBN 978-9949-19-369-1 (PDF)

Autoriõigus: Jüri Reimand, 2010

Tartu Ülikooli Kirjastus
<http://www.tyk.ee/>
Tellimus nr 270

CONTENTS

List of Original Publications	6
Abstract	8
Biological prerequisites	9
Principles of gene regulation	9
Methods of high-throughput biology	10
Complex networks in biology	10
Functional knowledge and analysis	11
Computational prerequisites	12
Enrichment analysis	12
Binomial probability	12
Hypergeometric probability	12
Multinomial probability	13
Multiple testing correction	13
Linear and logistic regression	14
Linear regression	14
Maximum likelihood	15
Analysis of Variance	16
Stepwise modelselection	17
eGLM: Ensemble Generalised Linear Models	19
eGLM algorithm	19
eGLM application on the gene regulation of yeast cell cycle	21
Summary of publications	25
Functional analysis of gene lists (Publication <i>I</i>)	25
Functional analysis of gene networks (Publication <i>II</i>)	26
Pathway animations of high-throughput data (Publication <i>III</i>)	27
Regulator knockout microarray analysis (Publication <i>IV</i>)	28
eGLM predicts novel regulators of quiescence (Publication <i>V</i>)	29
Conclusion	31
Acknowledgements	43
Summary in Estonian	44

LIST OF ORIGINAL PUBLICATIONS

Publications included here

1. 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* (2007) 35:W193–200.
2. Reimand J¹, Tooming L¹, Peterson H, Adler P, Vilo J: **GraphWeb: mining heterogeneous biological networks for gene modules with functional significance**. *Nucleic Acids Res* (2008) 36:W452–459.
3. Adler P¹, Reimand J¹, Janes J, Kolde R, Peterson H, Vilo J: **KEGGanim: pathway animations for high-throughput data**. *Bioinformatics* (2008) 24:588–590.
4. Reimand J, Vaquerizas JM, Todd AE, Vilo J, Luscombe NM: **Comprehensive reanalysis of transcription factor knockout expression data in *Saccharomyces cerevisiae* reveals many new targets**². *Nucleic Acids Res* (2010) Advance Access published online on April 12 2010, doi:10.1093/nar/gkq232.
5. Reimand J, Aun A, Vaquerizas JM, Todd AE, Bourgon R, Vilo J, Sedman J, Luscombe NM: **Ensemble models of perturbation, nucleosome positioning and protein-DNA interactions reveal novel G0 regulators in *Saccharomyces cerevisiae***. In preparation.

¹Equal contribution

²Included supplementary material was originally published online in *Nucleic Acids Res*

Other publications

1. Adler P¹, Peterson H¹, Agius P, Reimand J, Vilo J: **Ranking genes by their co-expression to subsets of pathway members.** *Ann NY Acad Sci* (2009) 1158:1–13.
2. Krushevskaya D, Peterson H, Reimand J, Kull M, Vilo J: **VisHiC – hierarchical functional enrichment analysis of microarray data.** *Nucleic Acids Res* (2009), 37:W587–592.
3. Adler P¹, Kolde R¹, Kull M, Tkachenko A, Peterson H, Reimand J, Vilo J: **Mining for coexpression across hundreds of datasets using novel rank aggregation and visualization methods.** *Genome Biol* (2009), 10:R139.
4. Schulz H, Kolde R, Adler P, Aksoy I, Anastassiadis K, Bader M, Billon N, Boeuf H, Bourillot PY, Buchholz F, Dani C, Doss MX, Forrester L, Gitton M, Henrique D, Hescheler J, Himmelbauer H, Hubner N, Karantzali E, Kretsovali A, Lubitz S, Pradier L, Rai M, Reimand J, Rolletschek A, Sachinidis A, Savatier P, Stewart F, Storm MP, Trouillas M, Vilo J, Welham MJ, Winkler J, Wobus AM, Hatzopoulos AK: **The FunGenES database: a genomics resource for mouse embryonic stem cell differentiation.** *PLoS ONE* (2009), 4:e6804.

¹Equal contribution

ABSTRACT

Modern high-throughput technologies provide scientists with large quantities of data that characterise life at the molecular level. These include expression levels of thousands of genes, regulatory DNA patterns, and complex networks of interacting molecules, among others. Analysis and interpretation of these data is an important task and comprises computational challenges.

This dissertation includes methods, computational experiments and software involving functional analysis of gene lists, networks and regulatory systems. Functional analysis is an interpretative task that associates experimental findings with existing knowledge of biological processes, functions, pathways, regulatory features, etc. The knowledge is often incomplete and statistical enrichment analysis helps generate hypotheses of underlying regulation and function. As part of this dissertation, three web-based tools were developed and published, including g:Profiler, GraphWeb and KEGGanim. The tools perform functional analysis of gene lists, interaction networks and biological pathways.

The concept of enrichment was further developed to logistic regression models that allow integration of complex biological knowledge and regulatory features in a more sensitive way. This dissertation involves the development of the eGLM (Ensemble Generalised Linear Models) algorithm that performs feature selection in categorical data. eGLM highlights predictors that best classify response in a logistic regression setting. Rather than fitting a potentially large set of predictors simultaneously, eGLM aggregates predictor scores from a large ensemble of small models. In the context of gene regulation, eGLM can be applied to the inference of process-specific gene regulators from biological measurements.

The biological contribution of the dissertation is the functional analysis of the transcription regulatory network of an important model organism, the yeast *Saccharomyces cerevisiae*. A high-confidence compendium of regulator profiles was compiled from multiple lines of high-throughput regulatory evidence and validated with functional analysis. The eGLM method was then applied to the compendium for computational prediction of process-specific transcription factors. eGLM showed high performance on the well-described cell cycle pathway and outperformed several common methods, as it recovered all core regulators with low error rate. Analysis of the cryptic state of quiescence (G_0) revealed functions like ageing and metabolic decline that agree with current knowledge. Furthermore, experimental validation demonstrates the power of eGLM, as several deletion mutants of predicted regulators have significantly different survival profiles in G_0 .

BIOLOGICAL PREREQUISITES

Principles of gene regulation

In its broadest sense, eukaryotic gene expression is the production of proteins using the information encoded in the DNA sequence. The first stage of gene expression termed transcription occurs in the nucleus and involves copying of gene sequences in the DNA as messenger RNA (mRNA). mRNA is then transported out of the nucleus to the ribosomes. During translation, amino acids are sequentially assembled into proteins based on mRNA code. Gene regulation determines the quantitative, spatial and temporal directives for gene expression.

In the context of this work, details of transcriptional regulation are of most importance. This stage of gene regulation is largely carried out by specific proteins called transcription factors (TFs). TFs recognise specific short DNA patterns termed transcription factor binding sites (TFBS) in so-called promoter areas upstream of gene start sites. TFs initiate transcription by attracting the RNA polymerase complex.

Transcriptional regulation is closely coupled with epigenetic mechanisms that govern DNA chromatin structure. The DNA molecule is packaged around specific protein complexes termed nucleosomes in a similar manner to beads on a string. Chromatin modifiers (CMs) are responsible for modifications in this structure. Denser DNA packaging provides less access to TFs and generally causes repression, while looser packaging allows access for transcriptional regulation.

Such a model is incomplete and additional regulatory mechanisms occur at every step. Transcriptional regulation can be carried out by multiple collaborating or competing TFs and CMs in activating or silencing manner. Special classes of RNAs are never translated into proteins and perform regulatory tasks, for instance microRNAs modify gene expression by labelling specific mRNAs for degradation. Gene splicing selects distinct mRNA regions for translation and discards others. The final role and lifetime of proteins is governed by post-translational folding and modifications. Some regulatory programs differ substantially across tissues, species and disease states, while others are conserved from bacteria to mammals. Biological function can be only rarely attributed to single genes and proteins and regulatory mechanisms emerge as complex networks.

Gene regulation is an active area of research. While some systems are relatively well described, most regulatory circuits remain uncharacterised. Models of common knowledge are constantly tested with evidence of novel mechanisms.

Methods of high-throughput biology

Traditional molecular biology has often focused on single genes or systems of few components. High-throughput technologies have changed this rhythm of research. For instance, microarrays provide simultaneous genome-wide measurements of mRNA expression [Schena1995, Spellman1998]. Since the very first experiments, microarrays have been used in a variety of applications, such as detecting cancer-related genes by comparing healthy and diseased tissues [DeRisi1996].

Microarrays can be combined with chromatin immunoprecipitation (ChIP-chip) to describe genome-scale TF binding maps and target genes [Ren2000]. Prominent examples of ChIP-chip application include analysis of self-renewal and differentiation in embryonic stem cells [Boyer2005]. Experimental assessment of protein-DNA interactions is closely coupled with *in silico* algorithms that predict TFBS from DNA sequence [Vilo2002, Wasserman2004]. Microarray-based techniques have application in many other tasks, including analysis of single nucleotide polymorphisms and copy number variations, binding sites of CMs, DNA methylation patterns, etc. Recent developments in next-generation sequencing lead the way for high-resolution technologies, including quantification of gene expression and TFBS [Metzker2010].

Gene perturbation is a powerful technique, as specific phenotypic changes give insight into gene function and hierarchy in a systematic way [Avery1992]. Microarrays provide enhanced resolution of genome-wide expression patterns and successfully scale to existing methods of phenotypic inference [VanDriessche2005]. Gene expression analysis of TF knockout mutants is particularly informative of gene regulation, as differential expression patterns reflect networks of direct and indirect regulatory targets [Hu2007, Reimand2010].

Complex networks in biology

The appearance of prominent network theoretical studies of computational biology coincides with the advent of high-throughput technologies. Investigators revealed curious structural properties that are common to regulatory networks of metabolic, transcriptional, and proteomic entities of the cell [Jeong2000, Oltvai2002, Maslov2002]. Complex biological networks are characterised by log-linear degree distribution and small-world property, meaning that most genes (proteins, metabolic compounds, etc) are linked via only a few edges. This global connectivity is provided by a small number of high-degree nodes, often corresponding to molecules that are critical for survival [Jeong2001]. Modularity and hierarchical structure is reflected in relatively separated subnetworks (topological modules) that may carry out distinct biological functions [Yook2004].

Several types of high-throughput biological data naturally appear as networks, and graph-based methods are increasingly used for interpretation [Carter2005, Aittokallio2006]. Protein-protein interaction networks comprise molecules that form stable complexes or establish transient contacts to carry out signal transduction

[Uetz2000, Collins2007, Yu2008]. Recent technological advancements in high-throughput genetic interaction screening quantify genome-wide associations by comparing the impact of single and double deletions [Tong2001, Costanzo2010]. ChIP-chip measurements, TFBS maps and regulator perturbation microarrays can be represented as regulatory networks of transcription factors and target genes [Harbison2004, MacIsaac2006, Hu2007, Reimand2010]. Interpretation of such networks is hindered by complexity, as global network connectivity establishes causal relations between most or all genes and proteins.

Functional knowledge and analysis

Data abundance is characteristic to all high-throughput biotechnologies. Statistical analysis of experimental results such as mRNA expression routinely leads to hundreds of genes that are significant to the initial hypothesis. Biological networks introduce additional complexity through the combinatorics of interactions. Interpretation of large-scale results is an important challenge.

Functional analysis is defined here as an interpretative task that uses existing knowledge to associate meaning to co-occurring genes [Huang2009b]. Since the body of biological knowledge is incomplete, statistical analysis helps identify enriched functional annotations that appear more frequently in the list than expected by random chance. The goal of enrichment analysis is twofold: first, the specifics of experimental design and enriched gene annotations may provide clues about underlying regulation, and second, one may infer the function of other, previously unannotated genes through evidence of co-occurrence in gene lists.

Various structured databases store biological knowledge in a systematic way. Gene Ontology (GO) contains broad and specific descriptions of molecular functions, cell localisation, and biological processes, and the scientific community constantly updates GO annotations of genes and proteins [Ashburner2000]. Biological pathways comprise small-scale systems of genes, proteins, and metabolic compounds that work together to carry out common function and often share features across the tree of life. Pathway databases like KEGG, Reactome, MetaCyc and others contain detailed interaction maps that sometimes reflect decades of research [Kanehisa2010, Vastrik2008, Caspi2006]. For instance, KEGG stores metabolic pathways, genetic and environmental signalling pathways, cellular and organismal processes, as well as human diseases and pathways for drug discovery. Interpretation and extension of pathways is an active area of research. For instance, my colleagues and I are developing co-expression based methods for pathway reconstruction with microarray data [Adler2009, Adler2009b].

Low-level functional evidence can be also applied in enrichment analysis. As an example, computationally predicted microRNA target sites and TFBS provide clues about alternative regulatory mechanisms [GriffithsJones2008, Matys2006]. First microarray studies involved clustering of expression profiles and search for enriched TFBS in corresponding gene promoters [Eisen1998, Brazma1998], and many advanced methods have emerged more recently [Troyanskaya2005].

COMPUTATIONAL PREREQUISITES

Enrichment analysis

The subject of enrichment analysis is a finite sample of categorical objects, each of which belongs to a class. Its goal is to compare observed class frequencies with theoretical, expected frequencies. A sample is said to be enriched in a class if objects of that class occur more frequently than expected by random chance; the class is said to be depleted if this proportion is less than expected. Statistical methods are used to quantify enrichment and depletion.

In the context of this work, statistical enrichment is used to evaluate the abundance of gene annotations in functional analysis. Given a list of genes from an experiment or a topological module in an interaction network, known biological processes, functions and pathways are considered in statistical enrichment tests.

Binomial probability

The simplest and most general enrichment test is based on the binomial distribution and applies to binary classification, e.g., covering successes and failures. If p is the expected probability of observing a success, n is the number of objects in the sample and k is the number of objects of successes in the sample, then the binomial probability $\mathbf{P}[X = k]$ reflects the chance of randomly observing exactly $X = k$ successes. $\mathbf{P}$ is computed as

$$\mathbf{P}[X = k] = \frac{n!}{k! \cdot (n - k)!} \cdot p^k \cdot (1 - p)^{n-k} = \binom{n}{k} \cdot p^k \cdot (1 - p)^{n-k}.$$

Note that the expected probability of observing a success is constant in the series, meaning that objects are sampled with replacement. As an example, one may compute the probability of observing $k = 60$ tails in a series of $n = 100$ tosses with a fair coin ($p_{\text{tail}} = 0.5$).

Hypergeometric probability

Hypergeometric distribution and corresponding test follow a similar concept as above, but account for successes without replacement. In other words, the probability of observing a success will change depending on how many successes are left in a finite collection of objects (that is, universe). If n is the number of objects in the sample, k is the number of successes in the sample, K is the total number of successes in the universe, and N is the total number of objects in the universe, then the hypergeometric probability $\mathbf{P}[X = k]$ reflects the chance of observing

exactly $X = k$ successes randomly. $\mathbf{P}$ is computed as

$$\mathbf{P}[X = k] = \frac{\binom{K}{k} \cdot \binom{N-K}{n-k}}{\binom{N}{n}}.$$

As an example, one may compute the probability of drawing exactly $k = 4$ white balls out of $n = 10$ balls from an urn with $N = 100$ black and $K = 5$ white balls.

Binomial and hypergeometric probabilities reflect exact samples of successes and failures and cover only a small fraction of potential configurations. In order to account for enrichment or depletion as a p-value, one needs to consider the observed sample together with other, more extreme samples. The cumulative hypergeometric p-value for the enrichment $\mathbf{P}$ is computed as a sum as

$$\mathbf{P}[X \geq k] = \sum_{x=k}^n \frac{\binom{K}{x} \cdot \binom{N-K}{n-x}}{\binom{N}{n}}.$$

Alternatively, one may test for depletion of successes with $x \in \{0, \dots, k\}$.

The values of x determine the meaning of binomial and hypergeometric p-values. In case of enrichment, one assumes an unexpectedly large proportion of successes and accounts for larger and more extreme proportions. In the opposite in case of depletion, as one assumes less than expected successes in the sample and accounts for smaller, more extreme configurations.

Multinomial probability

Binomial and hypergeometric distributions account for binary classification. Multinomial distribution and test comprise an extension to the binomial method for multiple classes. The multinomial probability $\mathbf{P}$ reflects the chance of randomly observing exactly $X_1 = x_1, \dots, X_k = x_k$ occurrences of each class, as

$$\mathbf{P}[X_1 = x_1, X_2 = x_2, \dots, X_k = x_k] = \frac{n!}{x_1! \cdot x_2! \cdot \dots \cdot x_k!} \cdot p_1^{x_1} \cdot p_2^{x_2} \cdot \dots \cdot p_k^{x_k},$$

Here, n stands for sample size, k is the number of potential classes, $x_1, \dots, x_k$ are the observed counts for these classes, and $p_1, \dots, p_k$ are expected probabilities.

Multiple testing correction

The concept of multiple comparisons is essential in statistical analysis if it involves simultaneous testing of several hypotheses. Due to the nature of statistical testing, inference errors and highly significant p-values are likely to occur more often if a collection of tests is considered as a whole. Multiple testing procedures account for the bias by reducing the significance all related p-values.

One of the most common multiple testing procedures is the False Discovery Rate (FDR) that controls the expected proportion of falsely rejected hypotheses in a series of independent tests [Benjamini1995]. If n is the number of tests, $H_1, \dots, H_n$ is a collection of alternative hypotheses and $\mathbf{P}_1, \dots, \mathbf{P}_n$ is a collection of p-values, then FDR ranks the p-values in increasing order $\mathbf{P}_{(1)}, \dots, \mathbf{P}_{(n)}$,

and for a given α , finds the largest k such that

$$\mathbf{P}_{(k)} \geq \frac{k}{n} \cdot \alpha.$$

All alternative hypotheses $H_{(1)}, \dots, H_{(k)}$ are then accepted as significant. The significance level α determines the fraction of expected false positives among corrected results, for instance $\alpha = 0.05$ suggests that 95% of significant p-values represent true positive alternative hypotheses.

Linear and logistic regression

Linear regression

Regression analysis determines how a normally distributed response variable Y is related to one or more predictor variables $X_1, \dots, X_k$ [Berthold2003, Venables2002]. Regression comprises a collection of n objects with values for predictor and response variables. The most common regression model is linear

$$Y \sim \beta_0 + \beta_1 X_1 + \dots + \beta_k X_k + \varepsilon,$$

where $Y = y_1, \dots, y_n$ is the response vector, $X_i = x_{1,i}, \dots, x_{n,i}, i \in \{1, \dots, k\}$ represent predictor vectors, $\beta = \beta_1, \dots, \beta_k$ stand for regression coefficients and β_0 is the baseline value of response (intercept), and

$$\mathbf{X} = \begin{pmatrix} x_{1,1} & \dots & x_{1,k} \\ \dots & \ddots & \dots \\ x_{n,1} & \dots & x_{n,k} \end{pmatrix}$$

is the predictor matrix. The vector $\varepsilon = \varepsilon_1, \dots, \varepsilon_n$ represents independent and normally distributed noise $\mathbf{N}(0, \sigma^2)$ that captures all influences to the response that are not reflected by the predictor variables. The expected value of response $\mathbf{E}(Y)$ is a β -weighted sum of predictor variables X_i :

$$\mathbf{E}(Y) = \beta_0 + \sum_{i=1}^k \beta_i X_i.$$

Initially unknown coefficients $\beta = \beta_1, \dots, \beta_k$ define the slopes of the regression line. These are inferred from the data as $\hat{\beta}$, such that the sum of distances between predictor variables and the response variable is minimal:

$$\hat{\beta} = \underset{\beta}{\operatorname{argmin}} ||Y - \mathbf{X}\beta||^2.$$

Linear regression coefficients can be computed explicitly using the Ordinary Least Squares (OLS) algorithm:

$$\hat{\beta} = (\mathbf{X}'\mathbf{X})^{-1} \mathbf{X}'\mathbf{y} = \frac{1}{n} \left(\sum_{i=1}^n x_i x_i' \right)^{-1} \frac{1}{n} \sum_{i=1}^n x_i y_i.$$

Maximum likelihood

Given a particular distribution and its known parameters, one may compute the probability of observing a fixed sample. More often, true parameters are unknown and only represented via a particular sample. One needs to infer the parameters that best describe the sample.

If $Z = z_1, \dots, z_n$ make up a sample of independent observations from a probability density $f(Z|\theta)$ with parameter vector θ , then $\mathbf{L}[\theta|Z]$ measures the relative likelihood that the density of θ has produced the observed sample as

$$\mathbf{L}[\theta|Z] = \prod_{i=1}^n f(z_i|\theta).$$

The equivalent log-likelihood function $\mathbf{l}$ is computationally more efficient:

$$\mathbf{l}[\theta|Z] = \ln \mathbf{L}[\theta|Z] = \sum_{i=1}^n \ln f(z_i|\theta).$$

Maximum Likelihood Estimation (MLE) is the process of finding such a parameter vector $\hat{\theta}$ that maximises likelihood $\mathbf{L}$:

$$\hat{\theta} = \hat{\mathbf{L}}[\theta|Z] = \underset{\theta}{\operatorname{argmax}} \mathbf{L}[\theta|Z].$$

MLE finds a particular set of parameters $\hat{\theta}$ so that the observed sample Z is more likely to have come from $f(Z|\hat{\theta})$ than from $f(Z|\theta)$ for any other values of θ . MLE also produces the estimated MLE value that corresponds to the optimal parameters, denoted here as $\hat{\mathbf{L}}_e$. In practice, the log variant $\hat{\mathbf{l}}_e$ of the above is used.

In the context of linear models, parameter estimation involves the selection of coefficients $\hat{\beta}$ that give the best linear relation between predictor variables $\mathbf{X}$ and the response variable Y . MLE of regression coefficients β is equivalent to corresponding derivation via OLS [Berthold2003]:

$$\hat{\beta} = (\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'y = \hat{\mathbf{L}}(\theta|Y, \mathbf{X}).$$

Generalised linear regression and logistic regression

Generalised linear models (GLM) provide two important extensions to linear regression. First, the response Y does not need to be normally distributed and can belong to any distribution of the exponential family. Second, predictors and expected response need not be linked linearly, but the association can be mapped to the linear space via a link function g , as

$$g(\mathbf{E}(Y)) = \beta_0 + \sum_{i=1}^k \beta_i X_i.$$

This generalisation is of great interest from the perspective of enrichment analysis, since linking permits the use of categorical data. Here objects represent genes, while predictors and response are categorical representations of functional annotations or regulatory elements.

If the predictor variables $X_1, \dots, X_k$ take one of two categorical values, these are applied as numerical indicator variables (either 0 or 1) that indicate the occurrence of one category (success). If a predictor X_i takes one of $c > 2$ categorical values, an extended model reflects additional categories as

$$Y \sim \beta_0 + \beta_1 X_1 + \dots + \beta_{i,1} X_{i,1} + \dots + \beta_{i,c-1} X_{i,c-1} + \dots + \beta_k X_k + \varepsilon,$$

where $\beta_{i,c}$ is implicitly defined via $\beta_{i,1}, \dots, \beta_{i,c-1}$ and omitted.

Binomial logistic regression applies if the response variable Y can take two values, e.g., success and failure, as

$$g(\mathbf{E}(Y)) = \log \frac{\mathbf{P}}{1 - \mathbf{P}} = \beta_0 + \sum_{i=1}^k \beta_i X_i.$$

Here $\mathbf{P}$ is the binomial probability of observing a successful response within objects with a specific combination of predictor indicators. The link g is also known as the logit function. Binomial logistic regression accounts for the probability of success in the response through the additive effects of indicator variables.

Multinomial logistic regression applies when the response variable Y takes more than two values. In that case, the logit $\mathbf{P}$ corresponds to the multinomial probability of observing the exact configuration of response categories within a specific combination of predictor indicators. Intuitively, logistic regression models the relationship between predictors and response by highlighting unexpectedly large or small proportions in combinations of categories.

Logistic regression coefficients β are estimated from the data using MLE. However, OLS is not directly applicable and advanced optimisation-based approaches such as Iteratively Reweighted Least Squares are needed [Holland1977].

Analysis of Variance

Analysis of Variance (ANOVA) is the procedure of evaluating the contributions of predictors $X_1, \dots, X_k$ in a fitted model, reflected in the estimated regression coefficients $\hat{\beta}_1, \dots, \hat{\beta}_k$. The interpretation of evaluation is twofold. First, $\hat{\beta}_i$ that are sufficiently close to zero have little contribution to the response, and corresponding predictors can be removed from the model. Second, $\hat{\beta}_i$ that produce a significant change in model variance upon removal correspond to important predictors that can be ranked according to significance.

One can define the fitted values $\hat{Y}_j = \hat{y}_1, \dots, \hat{y}_n$ that the model would predict as response given the existing predictor values X_i and coefficient estimates $\hat{\beta}_i$, as

$$\hat{Y} = \hat{\beta}_0 + \sum_{i=1}^k \beta_i X_i.$$

The residuals $R = Y - \hat{Y}$ quantify the difference between fitted and observed values in a linear model and relate to σ^2 , the variance in the noise vector ε , as

$$S^2 = \frac{\sum_{j=1}^n (y_j - \hat{y}_j)^2}{n - k + 1} = \frac{RSS}{v}.$$

Here, RSS stands for residual sum of squares and v represents residual degrees of freedom. RSS is interpreted as the general goodness of fit between model and data, while v quantifies model complexity through sample size and fitted parameters. ANOVA for linear models applies RSS for predictor evaluation. ANOVA for logistic regression is applied in a similar manner, but the deviance measure $\mathbf{D}$ with MLE is used for goodness of fit estimate instead.

If a given $\hat{\beta}_i$ is sufficiently close to zero and its contribution to response Y is insignificant, the goodness of fit of the model should not change significantly if X_i is discarded. Let $\mathbf{m}$ be an initial model with all coefficients $\beta_1, \dots, \beta_k$, and $\mathbf{m}_{\setminus i}$ be a model lacking a predictor variable $X_i, i \in \{1, \dots, k\}$. The deviance measure $\mathbf{D}$ is used to assess the contribution of the predictor X_i , computed as the difference of log MLE of the two models:

$$\mathbf{D}(\mathbf{m}, \mathbf{m}_{\setminus i}) = -2 \cdot (\hat{\mathbf{l}}_e(\mathbf{m}_{\setminus i}) - \hat{\mathbf{l}}_e(\mathbf{m})).$$

The significance of $\mathbf{D}$ is quantified approximately with the χ_v^2 distribution. If $z_1, \dots, z_v$ are independent random variables, then the sum

$$\chi_v^2 \sim \sum_{i=1}^v z_i^2$$

corresponds to a χ_v^2 distribution with v degrees of freedom. In case of logistic regression, degrees of freedom v can be interpreted as model complexity $\mathbf{C}$, corresponding to the number of model parameters as pairwise combinations of predictor and response classes. The significance of deviance is computed as

$$\mathbf{P}[\mathbf{D}(\mathbf{m}, \mathbf{m}_{\setminus i})] = \chi_{\mathbf{C}(\mathbf{m}) - \mathbf{C}(\mathbf{m}_{\setminus i})}^2 \mathbf{D}(\mathbf{m}, \mathbf{m}_{\setminus i}).$$

Accounting for model complexity is essential in regression analysis. Increasing complexity leads to overfitting, as any effects in the data can be explained with little error given a sufficient amount of parameters.

Stepwise model selection

Optimal model selection is a complex problem due to the combinatorial nature and extent of the potential model space. Several standard approaches exist for regression models, for instance the greedy step-wise selection (STEP) performs multiple comparisons of models by adding and/or removing predictors [Venables2002]. Each step is taken towards the direction that maximises the gain in fit as quantified by ANOVA. Backward selection attempts to simplify a given model by removing predictors in turn, while forward selection proposes additional predictors from an extended set. Advanced approaches alternate between the two directions, and allow search in higher-level terms, e.g., two-way predictor interactions. Stepwise selection generally terminates when no further steps improve the fit, or there are no more predictors to remove or add.

The goal of selection is to determine an optimal set of predictors that contribute significantly to the response and discard others that have no influence on predictions. In the gene regulatory context of this work, model selection aims to

detect significant gene regulators that classify process-specific genes. This model assumes that probabilities of process-specific genes are in logistic relation to combinations of regulatory features.

In logistic regression models, stepwise procedures use Akaike Information Criterion (AIC) for model selection [Akaike1974]. AIC is similar to deviance as it combines goodness of fit with MLE and complexity $\mathbf{C}$ with model parameter count. However, it does not compare related models, but provides a measure for a single model $\mathbf{m}$ as

$$\mathbf{A}(\mathbf{m}) = 2 \cdot \mathbf{C}(\mathbf{m}) - 2 \cdot \hat{\mathbf{I}}_e(\mathbf{m}).$$

eGLM: ENSEMBLE GENERALISED LINEAR MODELS

eGLM algorithm

In the following, I outline an algorithm based on logistic regression that performs categorical feature selection. The goal of the algorithm is to highlight significant predictors of response in a large collection of possible predictors. This method, termed Ensemble Generalised Linear Models (eGLM), aggregates predictions from a large number (an ensemble) of small models, rather than fitting all predictors simultaneously in a high-dimensional setting. In essence, eGLM is similar to the random forests classification method that grows a large number of decision trees [Breiman2001].

The eGLM method consists of two steps. The first step (Algorithm 1, GLM-step) performs logistic regression in a model of few predictors and evaluates their significance, using the methods described above. The second step (Algorithm 2, e-step) executes the GLM-step in a large-scale manner to create a model ensemble, and aggregates statistically significant predictor p-values as scores.

Algorithm 1 GLM-step: Fit a small model and evaluate predictors

Input: Response variable Y , predictor variables $X_1, \dots, X_k, k \geq 2$.

Output: P-values of predictor significance $P = p_1, \dots, p_k$.

1. Set predictor p-values to ones:

$$P \Leftarrow p_1, \dots, p_k = 1, \dots, 1.$$

2. Perform logistic binomial or multinomial regression:

$$\mathbf{m} \Leftarrow Y \sim \beta_0 + \varepsilon + \sum_{i=1}^k \beta_i X_i.$$

3. Perform backward selection with the STEP procedure:

$$\mathbf{m}' \Leftarrow \text{STEP}(\mathbf{m}).$$

4. Return P if $\mathbf{m}'$ contains less than 2 predictors.
5. For every predictor X_j of the optimised model $\mathbf{m}'$, perform ANOVA assessment and replace relevant p-value $p_j \in P$:

$$\forall j(Y_j \in \mathbf{m}') : p_j \Leftarrow \mathbf{P}[\mathbf{D}(\mathbf{m}', \mathbf{m}'_{\setminus j})].$$

6. Return P .
-

Algorithm 2 e-step: Construct a model ensemble and aggregate predictor scores

Input: response variable Y , predictor variables $X_1, \dots, X_h$, model size k , ensemble size e , significance cutoff α , $0 < \alpha < 1$, $k \geq 2$, $k \ll h$, $e \gg h$.

Output: Scores of predictor significance $S = s_1, \dots, s_h$.

1. Set collection of predictor p-values to empty set:

$$P \Leftarrow \emptyset.$$

2. Set predictor scores to zeroes:

$$S \Leftarrow s_1, \dots, s_h = 0, \dots, 0.$$

3. Repeat times of ensemble size, $i \in \{1, \dots, e\}$:

- (a) Sample k predictors with a uniform distribution:

$$X_{i,1}, \dots, X_{i,k} \Leftarrow \mathbf{U}(X_1, \dots, X_h).$$

- (b) Fit a model with sampled predictors, evaluate each predictor:

$$p_{i,1}, \dots, p_{i,k} \Leftarrow \text{GLM}(Y, X_{i,1}, \dots, X_{i,k}).$$

- (c) Collect predictor p-values:

$$P \Leftarrow \{P, p_{i,1}, \dots, p_{i,k}\}.$$

4. Perform multiple testing correction on all p-values from the ensemble:

$$P \Leftarrow \text{FDR}(P).$$

5. Repeat scoring for every predictor $X_1, \dots, X_h$:

$$\forall j \in \{1, \dots, h\} : s_j \Leftarrow \sum_{\substack{p_{i,j} \in P \\ p_{i,j} \leq \alpha}} -\log_{10}(p_{i,j}).$$

6. Return S .
-

eGLM application on the gene regulation of yeast cell cycle

In this study, models of transcriptional regulation were composed in the following way. All 6253 yeast genes were interpreted as objects with 285 regulator profiles R as predictor variables. Each regulator profile corresponded to a single regulator and provided categorical descriptions to gene-regulator associations.

Three types of high-throughput data were covered: gene differential expression in regulator knockout mutants (*up*, *down*, *none*) [Hu2007, Reimand2010], presence of regulator binding site in gene promoter [Harbison2004, MacIsaac2006, Erb2006, Zhu2009], combined with nucleosome depletion of regulator binding sites from [Kaplan2008] as (*nTFBS*, *TFBS*, *none*). These datasets were integrated in a combinatorial manner into regulator profiles R , providing 9 distinct categories (all above plus *up:TFBS*, *down:TFBS*, *up:nTFBS*, *down:nTFBS*). The resulting predictor matrix is sparse, since only high-confidence regulatory features are used and the majority of data points are characterised by lack of signal (*none*, background factor). For further details on data processing and discretisation, see the methods sections of Publications IV and V.

The response variable C in this model corresponded to the cell cycle (CC) phase classification of all yeast genes. The majority of 6253 yeast genes were not CC-related and are therefore classified as the background set (*none*). The previously published high-confidence list of CC regulated genes comprised 395 genes [Lichtenberg2005]. In our model, these were factored in the order of biological occurrence ($G1 \leq S \leq G2 \leq M$). In short, a model with k regulator profiles can be written as

$$C \sim \beta_0 + \varepsilon + \sum_{i=1}^k \beta_i R_i,$$

where β_i stand for regression coefficients, ε for noise, and k for model size.

In order to assess the performance of the eGLM algorithm in different parameter settings, I designed and executed large-scale computational experiments to select CC regulators to CC genes. The yeast CC pathway is relatively well studied and comprises 9 TFs that regulate one another and downstream processes in a circular manner [Simon2001]. To evaluate model predictions, I used the Area Under Curve (AUC) criterion that is a balanced measure of sensitivity and specificity. AUC was used in a stringent setting of binary classification, where all regulators with non-zero scores were regarded as positives.

Each run comprised an ensemble of $e = 50000$ models and model sizes $k = \{2, 4, 6, 8\}$. I tested variations of eGLM, by disabling backward selection and disabling multiple testing correction. I integrated nucleosome positioning in several ways, by including only nucleosome-depleted sites, disregarding nucleosome depletion data, and incorporating subsets that reflect different biological conditions. I also performed the inference using gene expression data E and TF binding data B as independent predictors rather than integrating these into regu-

Exploring model space with ensembles

cell cycle pathway, FDR=0.005

ensembles sampled=100, total ensembles=50000

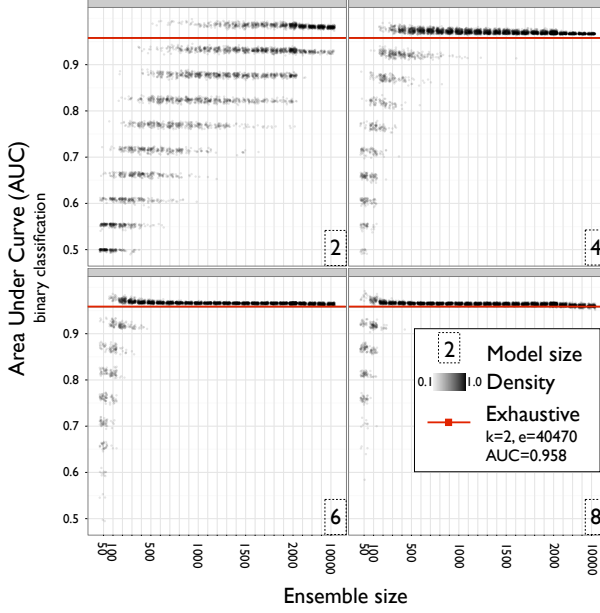


Figure 1. eGLM predictions depend on model space exploration via ensemble size e and model size k . Panels represent model sizes used in ensemble. X-axis stands for ensemble size (number of models) and Y-axis stands for Area Under Curve (AUC) in predicting cell cycle regulators. Every vertical point cloud represents 100 ensembles sampled out of 50000. Point color (grayscale) represents fraction of sampled ensembles with given AUC. Red line represents AUC for exhaustive pair-wise predictor ensemble with $e' = 40470$.

lator profiles. Models of this type can be written as

$$C \sim \beta_0 + \varepsilon + \sum_{i=1}^k \beta_i E_i + \gamma_i B_i,$$

where β_i and γ_i stand for regression coefficients.

The analysis provides several important conclusions about the eGLM algorithm. First, quality of predictions depends on ensemble size and model size (Figure 1). These two parameters reflect the proportion of model space and predictor combinations scanned during a run. Smaller ensembles naturally provide greater variance of AUC and less stability in predictions, as the covered model space is smaller. Consequently, less smaller ensembles may be used when every individual model comprises more predictors. The total number of predictors in the yeast regulatory system is relatively small and it is feasible to explore the whole space of pairwise predictor models exhaustively with

$$e' = \binom{h}{2} = \binom{285}{2} = 40470.$$

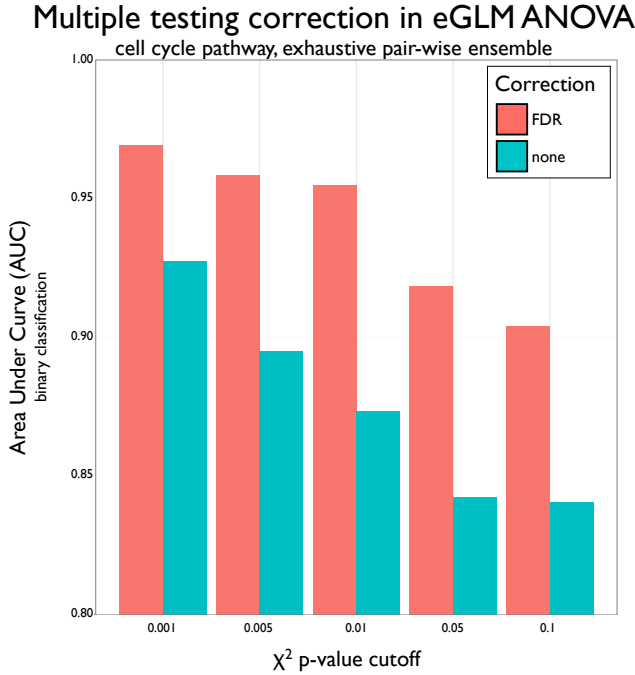


Figure 2. Global multiple testing correction improves predictions. X-axis represents p-value cutoff and Y-axis represents Area Under Curve (AUC) for cell cycle predictions (exhaustive pairwise models). Red bars represent predictions where all ensemble-wide χ^2 p-values of ANOVA tests were corrected with the FDR method, while blue bars represent predictions with no correction.

Moreover, exhaustive search appears to provide comparable AUC scores with ensembles that only approximately scan the model space. I have taken this approach here and in Publication V in order to gain high-confidence stable predictions. However, exhaustive analysis of more complex systems, e.g., mammalian transcriptional networks of thousands of regulators, is not computationally feasible and involves multiple testing issues. Therefore, approximate scanning should provide a viable alternative.

Second, quality of predictions justifies the steps taken in eGLM. Backward selection in individual models improves AUC, as it decreases the number of false positives (Publication V, Figure 5-D vs 5-B). Backward selection directly determines the subset of models for ANOVA scoring. Only models with at least two predictors can be used in ANOVA since it involves removal of predictors in turn. Predictions are also improved by global multiple testing correction (FDR) that is applied on the ensemble-wide collection of ANOVA χ^2 p-values (Figure 2). As FDR transforms p-values to more conservative levels, several false positive regulators fall below significance level and receive zero log-scores.

Third, integration of gene expression, TFBS, and nucleosome positioning data

as regulator profiles improves AUC and outperforms treatment as independent variables (Publication V, Figure 5-D vs 5-C). One likely explanation to this observation is the robustness and combinatorics characteristic to gene regulation [Kitano2004, Remenyi2004]. Collections of binding and expression target genes may represent partially redundant pathways. Combinatorial integration may help overcome the unexpectedly poor overlap between binding and expression measurements that has been reported previously [Hu2007, Balaji2008]. Incorporating nucleosome data with TFBS also improves predictions, as promoter areas of open chromatin are more likely targets of regulation (Publication V, Figures 5 and 6). This advantage has also been reported in recent work with protein-binding microarrays [Zhu2009, Gordan2009].

Last but not least, eGLM performs favourably in CC regulator predictions, in comparison to several other common statistical methods (Publication V, Figure 7). Notably, eGLM outperforms multinomial logistic regression with standard forward selection STEP that incrementally adds best predictors until convergence. Other comparisons, including ones with mutual information, hypergeometric p-value, and random forests, confirm the superiority of eGLM. The advantage of eGLM is especially evident in the stringent setting of binary classification where all predictors with non-zero scores are regarded as findings. Here I provide a statistically reasoned cut-off to distinguish significant predictors, while naïve application of common methods like mutual information only allows ranking. For further details on comparisons, see methods section of Publication V.

SUMMARY OF PUBLICATIONS

Functional analysis of gene lists (Publication I)

The main contribution of Publication I is the creation of g:Profiler, a web-based software for functional analysis of gene lists [Reimand2007]. A considerable number of Gene Ontology (GO) enrichment tools has been published during the last decade, notably DAVID [Huang2009a], Onto-Tools [Khatri2007], and Babelomics [AlShahrour2006] (see [Huang2009b] for a recent review). Several features distinguished g:Profiler at the time of publication. Tight links to the Ensembl database [Flicek2008] allow integration of numerous species, provide robust cross-linking between hundreds of kinds of IDs of proteins and genes, and establish ways to create new gene lists from public microarray databases and gene orthology mappings. In addition, g:Profiler includes two novel algorithms: efficient enrichment analysis of ordered lists, and selection of statistically significant results with multiple testing in the hierarchical structure of GO [Reimand2006].

ID mapping has proved to be a particularly useful feature with several applications in our research group and elsewhere. Integrated software includes GraphWeb, a tool for functional analysis of gene networks, Multi Experiment Matrix (MEM), a tool for co-expression analysis across large compendia of microarrays [Adler2009], KEGGanim, a pathway visualisation tool [Adler2008], and VisHiC, a microarray clustering and functional analysis tool [Krushevskaya2009]. The latter provides extends g:Profiler analysis – first, a given microarray dataset is processed with approximate hierarchical clustering [Kull2008], and second, biologically optimal clusters are selected with functional enrichment methods. Since the publication of g:Profiler, several novel features have been added, for instance analysis of gene lists originating from chromosomal regions, and g:Cocoa, a interface for enrichment analysis of multiple gene lists.

As an example of functional analysis, I studied microarrays of transcription factor deletion mutants of the yeast *S. cerevisiae* [Reimand2010]. Differentially expressed genes of the TF deletion mutant are likely targets of the deleted regulator in wild-type cells; if such targets comprise a large proportion of genes with specific annotations like *ribosome biogenesis and translation*, then the TF may be involved in ribosome regulation in a direct or indirect manner.

My role in the creation of g:Profiler is ubiquitous and covers most work with algorithms, data integration, and user interface design. I implemented algorithms for intersection and enrichment analysis on the basis of previous work on fast set operations [Vilo2002, Reimand2006]. The graphical representation of results was built using the SWOG environment [Hansen2005]. I also conducted simulations

to detect optimal significance cutoffs for GO, and was involved in the design of the analytical cutoff method. I implemented the majority of database connections and data integration tasks, including those of Ensembl, GO, Reactome, KEGG, and Gene Expression Omnibus. I designed the mechanisms for referencing gene IDs and orthologs based on the information retrieved from Ensembl and used the BerkeleyDB engine to create fast cross-linking indices. Last but not least, I had a major role in finalising and documenting the software and writing the manuscript. My colleagues integrated TFBS information from the Transfac database, created rapid similarity search algorithms for g:Sorter that were later published separately [Kull2008, Adler2009], and contributed significantly to the development of the analytical cutoff method.

Functional analysis of gene networks (Publication II)

The main contribution of Publication II is the creation of GraphWeb, a web-based tool for the integration and interpretation of complex biological networks [Reimand2008]. Functional analysis of genomic networks is an extension of the list-based methods implemented in g:Profiler [Reimand2007]. A number of tools of biological network analysis have been published, notably including Cytoscape, VisAnt, and SNOW [Cline2007, Hu2009, Minguéz2009]. The GraphWeb pipeline generally comprises two stages: first, one of several topological algorithms dissects the network into modules to reduce complexity, and second, detected modules are profiled with functional enrichment analysis. GraphWeb implements two types of topological algorithms: supervised algorithms extract predefined features with potential biological meaning; unsupervised algorithms use stochastic processes to distinguish denser areas (communities, modules) of the network. The supervised class of algorithms finds features like cliques, fully connected modules that have been likened to stable protein complexes in protein-protein interaction networks [Przulj2004], and hub-based modules of nodes and neighbours that would reflect TFs and target genes in a regulator perturbation network [Hu2007]. Unsupervised module detection algorithms select certain edges for disposal – Markov Clustering [Dongen2000] simulates stochastic flow to determine less-burdened edges, while betweenness centrality clustering finds edges that act as bottlenecks between communities. These algorithms have been used previously in many occasions, including protein complex detection and reconstruction of metabolic pathways [Krogan2006, Dunn2005]. GraphWeb provides further approaches to extend and filter biological networks, including support for directed and weighted edges, integration of multiple networks as layers, and integration of multi-species data via orthology maps.

As an example of GraphWeb application, I have analysed perturbation microarray data of *S. cerevisiae* to find interacting groups of proteins that are targeted in TF knockout mutants [Reimand2010]. To detect statistical enrichment of protein interaction modules for each TF mutant, I defined a set of 'core' connections comprising interactions among the differentially expressed genes, and a set of

'neighbourhood' connections that also include interactions between the core and non-differentially expressed genes. I then compared the number of interactions among core and neighbourhood genes, measuring the proportion and significance of interacting genes that are targeted by the same TF. Several regulators of different types appeared to target the same functional group of genes with GO enrichments in vitamin B6 metabolism, in an illustration to likely collaborative patterns in the regulation of this essential component of glycolysis.

My role in the creation of GraphWeb involves software and user interface design. My contribution is the concept of methods for functional analysis of gene networks via topological dissection and enrichment analysis. I selected the methods for dissection based on possible biological interpretation, and designed ways to extend networks, including integration of data via layers, multi-species networks via orthologs, and data filtering, among others. The majority of programming tasks, especially those related to network algorithms, were executed by the second equally-contributing author. Functional enrichment analysis as well as cross-referencing of gene IDs and orthologs is carried out by g:Profiler modules [Reimand2007]. My colleagues and I integrated the Markov Clustering implementation developed earlier [Dongen2000], the file format conversion methods developed in Cytoscape [Cline2007], and the topological algorithms of the GraphViz software [Gansner1999]. I had a major role in finalising the software user interface and writing the manuscript.

Pathway animations of high-throughput data (Publication *III*)

The main contribution of Publication *III* is the creation of KEGGanim, a web-based software for the visual interpretation of high-throughput data in the context of biological pathways [Adler2008]. Pathway databases as well as independent tools like PathwayExpress provide visualisation methods [Khatri2007, Kanehisa2010, Vastrik2008]. However at the time of publication, KEGGanim stood out for comprehensive animations. The input of KEGGanim is a matrix of gene expression values for one or more experimental conditions; the output is a KEGG pathway map where relevant proteins are coloured with values from the expression matrix. KEGGanim animations switch colours over time to reflect changes in gene expression in different experimental conditions. Animated time series of gene expression are particularly informative, as they provide visual evidence to consecutive molecular events.

As an example of KEGGanim application, my colleagues and I designed pathway snapshots of cell cycle and apoptosis using gene expression of mouse embryonic stem cell differentiation time course [Schulz2009]. These reveal complementary expression patterns of pluripotent and differentiated cells – fundamental processes like DNA replication and damage control become gradually down-regulated while pro-apoptotic genes are induced as differentiation progresses.

My role in creating KEGGanim involves finalisation as public web-based software, while methods for linking pathway maps and gene expression measure-

ments were mostly developed by the first equally-contributing author. I implemented methods for uploading, storing and processing microarray datasets, as well as simple protocols for user authentication, and integrated the previously developed nearest-neighbours algorithm [Troyanskaya2001, Montaner2006]. Cross-referencing of gene IDs is carried out by g:Profiler modules [Reimand2007]. I was involved in the design and implementation of the software user interface, and had a major role in preparing the manuscript.

Regulator knockout microarray analysis (Publication IV)

Several small-scale and large-scale perturbation microarray datasets have been created in the common genomics model organism budding yeast *Saccharomyces cerevisiae* [Svetlov1995, Hughes2000, Chua2006]. The newest dataset by Hu and colleagues is a compendium of 269 experiments that for the first time covers the vast majority of TFs of an eukaryotic organism [Hu2007]. As many TFs also have a known TFBS from ChIP-chip and more recent protein-binding microarrays [Harbison2004, Zhu2009], the dataset is of great relevance to computational studies of gene regulation. However, there are several fundamental drawbacks in the original preprocessing of the compendium that hinder use in follow-up studies. These include an outdated error model of differential expression, lack of microarray print-tip and background correction according to quality assessment tools [Vaquerizas2004], and lack of multiple testing procedures. Inadequate contrasts were applied in comparisons of normal and mutant cells with no consideration of secondary effects such as batches, strains and kits. Consequently, the authors and others have discussed poor overlap of the knockout compendium with other types of evidence, especially TFBS [Hu2007, Balaji2008].

The main contribution of Publication IV is the reprocessing of the Hu *et al* compendium [Hu2007] and consequent functional analysis to validate its improvements [Reimand2010]. In the first stage of the analysis, I followed state-of-the-art microarray processing procedures that have a wide consensus in the community and have been made public via the Bioconductor software suite [Nadon2002, Allison2006, Gentleman2004]. My pipeline included normalisation and corrections for print-tip and background with variance stabilisation [Huber2002], detection of differential expression and statistical significance with secondary effects using linear models and empirical Bayesian probabilities [Smyth2004], and false positive control with global multiple-testing correction [Benjamini1995]. I had multiple roles in this study, including design and execution of the microarray processing pipeline, methods for downstream functional analysis, biological interpretation of results, and manuscript preparation. My colleagues gave valuable advice on microarray analysis, provided high-confidence TFBS data and several examples of biological importance, and contributed significantly to the final manuscript.

The analysis resulted in a 10-fold increase in the number of differentially expressed genes and a 90% overlap with previous data, indicating the greater sen-

sitivity of the new approach and a vastly expanded repertoire of regulator targets. I performed various types of functional analyses, including GO and pathway enrichments, assessment of direct and indirect TFBS, and analysis of protein-protein interaction modules. All cases demonstrated the advantage of the new dataset.

Two results deserve further attention. First, as opposed to the previous analysis, the majority of deleted TFs showed significantly decreased levels in corresponding knockout mutants, which is expected as the factors have been deleted in the mutants. However, several regulators with the highest number of target genes lacked a significant depletion signal, probably because of the sensitivity threshold of the microarray platform. This strongly highlights the importance of sensitive microarray analysis, as even undetectably small changes in regulator expression may cause dramatic downstream events.

Second, the new analysis achieves a small but significant 4-fold improvement in the detection of TF target genes with corresponding TFBS in promoters. I extended this analysis to regulatory cascades of primary and secondary TFBS and looked for sites of regulators that are differentially expressed in response to the deleted TF. In that case, the advantage of the new approach becomes more evident – it explains more than 30% of all regulators and targets and provides a 25-fold improvement compared to the original analysis. This is probably an indication of tightly controlled circuits of TF-TF and TF-gene regulation.

eGLM predicts novel regulators of quiescence (Publication V)

Quiescence (stationary phase, G_0) is probably the most wide-spread and concurrently the least understood cellular state. It is characterised by the lack of proliferation, reduced metabolism, transcription and translation, silenced genomes, greater resilience to stress, and associations to a variety of important biomedical research topics, including ageing, cancer, neurodegenerative diseases, and microbiology [Gray2004]. The gene regulatory programs of G_0 are largely uncharacterised, but are believed to involve large-scale reprogramming of the cell.

The major contribution of Publication V (in preparation) is a computational inference study of regulatory programs and specific regulators in G_0 of *S. cerevisiae*. While numerous computational studies of gene regulation have been published since the advent of high-throughput biology [Tavazoie1999, Brazma1998, Chen1999, Alter2000, Liebermeister2002, Lee2003, Bussemaker2001, Beer2004, Margolin2006, Friedman2004, Segal2003, Peer2001, Rung2002, Markowitz2005], I have identified several arguments for the novelty of this work.

First, the scale and biological context of this work is highly relevant, as I compiled a high-confidence collection of gene regulatory features, including knockout profiles for most regulators in yeast [Hu2007, Reimand2010], and the most recent map of TFBS from several lines of evidence [Harbison2004, MacIsaac2006, Erb2006, Zhu2009]. Furthermore, TFBS data were integrated with another dimension of biological evidence: I distinguished sites with significantly lower nucleosome occupancy, meaning that these are located in regions of open chromatin

and are more likely targets of DNA-binding TFs [Kaplan2008]. After merging these datasets as regulator profiles, I made predictions for G_0 entry, maintenance and exit using public microarray collections [Radonjic2005, Aragon2008]. The regulatory setting of G_0 is important for fundamental questions of biology, and to my knowledge, no computational study of this extent has been yet conducted.

Second, several aspects of the computational method are novel, and benchmark predictions give promising results in the regulatory setting of the well-defined cell cycle pathway [Simon2001, Lichtenberg2005]. The details of the method, termed eGLM (Ensemble Generalised Linear Models), are detailed in this dissertation and involve standard procedures of logistic regression [Venables2002, Berthold2003]. The developed ensemble approach of aggregating predictions over many small models is similar to the random forests method [Breiman2001]. Using empirical evidence of recovering cell cycle TFs, I showed that eGLM is more successful than fitting all TFs in logistic regression, and moreover, it outperforms several other common methods. I had multiple roles in this study, including design and implementation of the eGLM method, computational experiments with cell cycle and G_0 genes, large-scale simulations to explore parameter space and intermediate eGLM variants, and downstream functional analysis and biological interpretation of predicted regulators and target genes. My colleagues gave valuable advice on linear models and data processing, helped interpret the biological significance of our findings, and conducted biological experiments with candidate regulators.

A great deal of this study is dedicated to the biological interpretation of predicted G_0 -specific regulators and target genes. The predictions are generally in good agreement with existing understanding of the quiescent state, as illustrated by the literature of transcriptional and epigenetic regulation, and functional annotations from GO and pathways. As an example, the pre-phase of G_0 in yeast is diauxic shift that involves nutrient depletion, switch to non-fermentable carbon sources and cease of proliferation [Gray2004]. In agreement with the above, I predict a number of G_0 entry regulators and target genes with broad types of functional evidence – cell respiration and response to starvation that reflect metabolic switch, and pheromone response with CC arrest reflecting decreased proliferation. Strikingly, a number of predicted TFs, interactions and target genes relate to regulatory programs of ageing, for instance including the silencing proteins SIR that have been related to restricted metabolic balance and prolonged lifespan in several model organisms [Guarente2000]. Another particularly interesting observation is the enrichment of genes related to methionine metabolism. This essential amino acid appears to have a regulatory role in the balance between prolonged life and providing for descendants, as a recent study showed that methionine-enriched but otherwise restricted diet of fruit flies restored normal egg-laying without decreasing prolonged lifespan [Grandison2009]. Last but not least, several TFs predicted in this work also were validated in biological experiments, as my colleagues and I showed that yeast cells lacking certain regulators are significantly more or less viable than wildtype cells in G_0 .

CONCLUSION

The topic of this dissertation is computational functional analysis of genomic data. It comprises methods, software and computational experiments that help interpret measurements of gene expression, protein-DNA interactions, complex interaction networks, and other types of evidence in the context of existing biological knowledge. This work can be broadly summarised as follows.

Three publicly available web-based tools were developed and published, including g:Profiler, GraphWeb and KEGGanim. g:Profiler permits analysis of one or more groups of genes, and provides additional features for building gene lists, orthology mapping, co-expression search and gene ID conversion. GraphWeb extends the analysis of gene lists to gene networks, by performing topological clustering of networks to find functional modules. KEGGanim links the visual representation of biological pathways with measurements of gene expression levels. These tools perform functional analysis using standard statistical enrichment tests detailed in the dissertation.

The functional enrichment concepts were further extended to regression models that allow better integration of complex categorical data such as gene regulatory features and functional annotations. Standard approaches of binomial and multinomial logistic regression were combined into an algorithm termed eGLM, for Ensemble Generalised Linear Models. eGLM scans the potentially large space of predictor variables using an ensemble of models of few predictors, rather than fitting a single model in a high-dimensional setting. The goal of eGLM is feature selection, that is, inference of predictors that best classify the response. In the gene regulatory setting, regulators to process-specific genes can be predicted.

The biological contribution of this work involves functional analysis of the yeast *S. cerevisiae*. A high-confidence dataset of genome-wide regulatory features was composed from public sources. This included reprocessing, functional analysis and re-publication of a transcription factor (TF) knockout microarray compendium. The latter was integrated with TF binding sites and evidence of nucleosome depletion. eGLM was then applied for predicting process-specific TFs. Its performance was assessed on the well-defined cell cycle pathway, core TFs were recovered with low error rate, and standard methods were outperformed. Next, the regulatory programs of quiescence (G_0) were inferred with eGLM. Functional analysis of predicted TFs and targets agreed with the understanding of G_0 and covered important functions such as ageing, metabolic decline, stress response, among others. Importantly, several predicted TFs validated in experimental setting, as corresponding mutants had significantly different survival profiles in G_0 .

The studies presented here provide multiple perspectives for further research.

In the context of functional knowledge, additional resources continue to appear as the field of computational biology matures. New databases and ontologies, as well as novel experimental measures and details can be added to the functional enrichment pipelines. Novel algorithms are required that allow functional analysis of additional data types, for instance weighted gene lists, dense gene networks, semi-cliques, overlapping network modules, etc. A development already in progress involves introduction of background gene lists to represent custom universes for g:Profiler analysis. Next-generation sequencing and novel dense microarrays introduce further challenges and novel semantics.

The reprocessed microarray compendium together with regulator profiles of expression, binding sites and nucleosome depletion data provide a useful starting point for genome-scale studies of gene regulation and resources of cross-validation for experimentalists and computational scientists alike. The proposed mechanisms and factors of transcriptional and epigenetic regulation in G_0 should be validated and explained in more elaborate experimental assays.

The eGLM method developed here can be applied on the condition of data availability in other important and more complex regulatory systems, such as mammalian self-renewal and differentiation circuits of embryonic stem cells. The algorithm may be applicable in other fields like computational finance to problems like predicting technical factors that contribute to significant changes in stock price movement.

BIBLIOGRAPHY

- [Adler2008] P. Adler, J. Reimand, J. Janes, R. Kolde, H. Peterson, and J. Vilo. KEGGanim: pathway animations for high-throughput data. *Bioinformatics*, 24:588–590, Feb 2008.
- [Adler2009] P. Adler, R. Kolde, M. Kull, A. Tkachenko, H. Peterson, J. Reimand, and J. Vilo. Mining for coexpression across hundreds of datasets using novel rank aggregation and visualization methods. *Genome Biol.*, 10:R139, 2009.
- [Adler2009b] P. Adler, H. Peterson, Agius P., J. Reimand, and J. Vilo. Ranking genes by their co-expression to subsets of pathway members. *Ann NY Acad Sci*, 1158:1–13, 2009.
- [Aittokallio2006] T. Aittokallio and B. Schwikowski. Graph-based methods for analysing networks in cell biology. *Brief. Bioinformatics*, 7:243–255, Sep 2006.
- [Akaike1974] H. Akaike. A new look at the statistical model identification. *Automatic Control, IEEE Transactions on*, 19(6):716–723, January 1974.
- [AlShahrour2006] F. Al-Shahrour, P. Minguéz, J. Tarraga, D. Montaner, E. Alloza, J. M. Vaquerizas, L. Conde, C. Blaschke, J. Vera, and J. Dopazo. BABELOMICS: a systems biology perspective in the functional annotation of genome-scale experiments. *Nucleic Acids Res.*, 34:W472–476, Jul 2006.
- [Allison2006] D. B. Allison, X. Cui, G. P. Page, and M. Sabripour. Microarray data analysis: from disarray to consolidation and consensus. *Nat. Rev. Genet.*, 7:55–65, Jan 2006.
- [Alter2000] O. Alter, P. O. Brown, and D. Botstein. Singular value decomposition for genome-wide expression data processing and modeling. *Proc. Natl. Acad. Sci. U.S.A.*, 97:10101–10106, Aug 2000.
- [Aragon2008] A. D. Aragon, A. L. Rodriguez, O. Meirelles, S. Roy, G. S. Davidson, P. H. Tapia, C. Allen, R. Joe, D. Benn, and M. Werner-Washburne. Characterization of differentiated quiescent and non-quiescent cells in yeast stationary-phase cultures. *Mol. Biol. Cell*, 19:1271–1280, Mar 2008.
- [Ashburner2000] M. Ashburner, C. A. Ball, J. A. Blake, D. Botstein, H. Butler, J. M. Cherry, A. P. Davis, K. Dolinski, S. S. Dwight, J. T. Eppig, M. A. Harris, D. P. Hill, L. Issel-Tarver, A. Kasarskis, S. Lewis,

- J. C. Matese, J. E. Richardson, M. Ringwald, G. M. Rubin, and G. Sherlock. Gene ontology: tool for the unification of biology. The Gene Ontology Consortium. *Nat. Genet.*, 25:25–29, May 2000.
- [Avery1992] L. Avery and S. Wasserman. Ordering gene function: the interpretation of epistasis in regulatory hierarchies. *Trends Genet.*, 8:312–316, Sep 1992.
- [Balaji2008] S. Balaji, L. M. Iyer, M. M. Babu, and L. Aravind. Comparison of transcription regulatory interactions inferred from high-throughput methods: what do they reveal? *Trends Genet.*, 24:319–323, Jul 2008.
- [Beer2004] M. A. Beer and S. Tavazoie. Predicting gene expression from sequence. *Cell*, 117:185–198, Apr 2004.
- [Benjamini1995] Y. Benjamini and Y. Hochberg. Controlling the false discovery rate: A practical and powerful approach to multiple testing. *Journal of the Royal Statistical Society. Series B (Methodological)*, 57(1):289–300, 1995.
- [Berthold2003] M. Berthold and DJ Hand. *Intelligent data analysis: an introduction*. Springer Verlag, 2003.
- [Boyer2005] L. A. Boyer, T. I. Lee, M. F. Cole, S. E. Johnstone, S. S. Levine, J. P. Zucker, M. G. Guenther, R. M. Kumar, H. L. Murray, R. G. Jenner, D. K. Gifford, D. A. Melton, R. Jaenisch, and R. A. Young. Core transcriptional regulatory circuitry in human embryonic stem cells. *Cell*, 122:947–956, Sep 2005.
- [Brazma1998] A. Brazma, I. Jonassen, J. Vilo, and E. Ukkonen. Predicting gene regulatory elements in silico on a genomic scale. *Genome Res.*, 8:1202–1215, Nov 1998.
- [Breiman2001] L. Breiman. Random forests. *Machine Learning*, 45(1):5–32, October 2001.
- [Bussemaker2001] H. J. Bussemaker, H. Li, and E. D. Siggia. Regulatory element detection using correlation with expression. *Nat. Genet.*, 27:167–171, Feb 2001.
- [Carter2005] G. W. Carter. Inferring network interactions within a cell. *Brief. Bioinformatics*, 6:380–389, Dec 2005.
- [Caspi2006] R. Caspi, H. Foerster, C. A. Fulcher, R. Hopkinson, J. Ingraham, P. Kaipa, M. Krummenacker, S. Paley, J. Pick, S. Y. Rhee, C. Tissier, P. Zhang, and P. D. Karp. MetaCyc: a multiorganism database of metabolic pathways and enzymes. *Nucleic Acids Res.*, 34:D511–516, Jan 2006.
- [Chen1999] T. Chen, H. L. He, and G. M. Church. Modeling gene expression with differential equations. *Pac Symp Biocomput*, pages 29–40, 1999.

- [Chua2006] G. Chua, Q. D. Morris, R. Sopko, M. D. Robinson, O. Ryan, E. T. Chan, B. J. Frey, B. J. Andrews, C. Boone, and T. R. Hughes. Identifying transcription factor functions and targets by phenotypic activation. *Proc. Natl. Acad. Sci. U.S.A.*, 103:12045–12050, Aug 2006.
- [Cline2007] M. S. Cline, M. Smoot, E. Cerami, A. Kuchinsky, N. Landys, C. Workman, R. Christmas, I. Avila-Campilo, M. Creech, B. Gross, K. Hanspers, R. Isserlin, R. Kelley, S. Killcoyne, S. Lotia, S. Maere, J. Morris, K. Ono, V. Pavlovic, A. R. Pico, A. Vailaya, P. L. Wang, A. Adler, B. R. Conklin, L. Hood, M. Kuiper, C. Sander, I. Schmulevich, B. Schwikowski, G. J. Warner, T. Ideker, and G. D. Bader. Integration of biological networks and gene expression data using Cytoscape. *Nat Protoc*, 2:2366–2382, 2007.
- [Collins2007] S. R. Collins, P. Kemmeren, X. C. Zhao, J. F. Greenblatt, F. Spencer, F. C. Holstege, J. S. Weissman, and N. J. Krogan. Toward a comprehensive atlas of the physical interactome of *Saccharomyces cerevisiae*. *Mol. Cell Proteomics*, 6:439–450, Mar 2007.
- [Costanzo2010] M. Costanzo, A. Baryshnikova, J. Bellay, Y. Kim, E. D. Spear, C. S. Sevier, H. Ding, J. L. Koh, K. Toufighi, S. Mostafavi, J. Prinz, R. P. St Onge, B. VanderSluis, T. Makhnevych, F. J. Vizeacoumar, S. Alizadeh, S. Bahr, R. L. Brost, Y. Chen, M. Cokol, R. Deshpande, Z. Li, Z. Y. Lin, W. Liang, M. Marback, J. Paw, B. J. San Luis, E. Shuteriqi, A. H. Tong, N. van Dyk, I. M. Wallace, J. A. Whitney, M. T. Weirauch, G. Zhong, H. Zhu, W. A. Houry, M. Brudno, S. Ragibzadeh, B. Papp, C. Pal, F. P. Roth, G. Giaever, C. Nislow, O. G. Troyanskaya, H. Bussey, G. D. Bader, A. C. Gingras, Q. D. Morris, P. M. Kim, C. A. Kaiser, C. L. Myers, B. J. Andrews, and C. Boone. The genetic landscape of a cell. *Science*, 327:425–431, Jan 2010.
- [DeRisi1996] J. DeRisi, L. Penland, P. O. Brown, M. L. Bittner, P. S. Meltzer, M. Ray, Y. Chen, Y. A. Su, and J. M. Trent. Use of a cDNA microarray to analyse gene expression patterns in human cancer. *Nat. Genet.*, 14:457–460, Dec 1996.
- [Dongen2000] S. van Dongen. *Graph Clustering by Flow Simulation*. PhD thesis, Centre for Mathematics and Computer Science, University of Utrecht, the Netherlands, 2000.
- [Dunn2005] R. Dunn, F. Dudbridge, and C. M. Sanderson. The use of edge-betweenness clustering to investigate biological function in protein interaction networks. *BMC Bioinformatics*, 6:39, 2005.
- [Eisen1998] M. B. Eisen, P. T. Spellman, P. O. Brown, and D. Botstein. Cluster analysis and display of genome-wide expression patterns. *Proc. Natl. Acad. Sci. U.S.A.*, 95:14863–14868, Dec 1998.

- [Erb2006] I. Erb and E. van Nimwegen. Statistical features of yeast's transcriptional regulatory code. *IEEE Proceedings ICCSB*, 2006.
- [Flicek2008] P. Flicek, B. L. Aken, K. Beal, B. Ballester, M. Caccamo, Y. Chen, L. Clarke, G. Coates, F. Cunningham, T. Cutts, T. Down, S. C. Dyer, T. Eyre, S. Fitzgerald, J. Fernandez-Banet, S. Graf, S. Haider, M. Hammond, R. Holland, K. L. Howe, K. Howe, N. Johnson, A. Jenkinson, A. Kahari, D. Keefe, F. Kokocinski, E. Kulesha, D. Lawson, I. Longden, K. Megy, P. Meidl, B. Overduin, A. Parker, B. Pritchard, A. Prlic, S. Rice, D. Rios, M. Schuster, I. Sealy, G. Slater, D. Smedley, G. Spudich, S. Trevanion, A. J. Vilella, J. Vogel, S. White, M. Wood, E. Birney, T. Cox, V. Curwen, R. Durbin, X. M. Fernandez-Suarez, J. Herrero, T. J. Hubbard, A. Kasprzyk, G. Proctor, J. Smith, A. Ureta-Vidal, and S. Searle. Ensembl 2008. *Nucleic Acids Res.*, 36:D707–714, Jan 2008.
- [Friedman2004] N. Friedman. Inferring cellular networks using probabilistic graphical models. *Science*, 303:799–805, Feb 2004.
- [Gansner1999] E.R. Gansner and S.C. North. An open graph visualization system and its applications to software engineering. *Software - Practice and Experience*, 30:1203–1233, 1999.
- [Gentleman2004] R. C. Gentleman, V. J. Carey, D. M. Bates, B. Bolstad, M. Dettling, S. Dudoit, B. Ellis, L. Gautier, Y. Ge, J. Gentry, K. Hornik, T. Hothorn, W. Huber, S. Iacus, R. Irizarry, F. Leisch, C. Li, M. Maechler, A. J. Rossini, G. Sawitzki, C. Smith, G. Smyth, L. Tierney, J. Y. Yang, and J. Zhang. Bioconductor: open software development for computational biology and bioinformatics. *Genome Biol.*, 5:R80, 2004.
- [Gordan2009] R. Gordan, A. J. Hartemink, and M. L. Bulyk. Distinguishing direct versus indirect transcription factor-DNA interactions. *Genome Res.*, 19:2090–2100, Nov 2009.
- [Grandison2009] R. C. Grandison, M. D. Piper, and L. Partridge. Amino-acid imbalance explains extension of lifespan by dietary restriction in *Drosophila*. *Nature*, 462:1061–1064, Dec 2009.
- [Gray2004] J. V. Gray, G. A. Petsko, G. C. Johnston, D. Ringe, R. A. Singer, and M. Werner-Washburne. "Sleeping beauty": quiescence in *Saccharomyces cerevisiae*. *Microbiol. Mol. Biol. Rev.*, 68:187–206, Jun 2004.
- [GriffithsJones2008] S. Griffiths-Jones, H. K. Saini, S. van Dongen, and A. J. Enright. miRBase: tools for microRNA genomics. *Nucleic Acids Res.*, 36:D154–158, Jan 2008.
- [Guarente2000] L. Guarente. Sir2 links chromatin silencing, metabolism, and aging. *Genes Dev.*, 14:1021–1026, May 2000.

- [Hansen2005] J. Hansen. Graafikakeel SWOG. Bachelor’s thesis (in estonian). Institute of Computer Science, University of Tartu, Estonia, 2005.
- [Harbison2004] C. T. Harbison, D. B. Gordon, T. I. Lee, N. J. Rinaldi, K. D. Macisaac, T. W. Danford, N. M. Hannett, J. B. Tagne, D. B. Reynolds, J. Yoo, E. G. Jennings, J. Zeitlinger, D. K. Pokholok, M. Kellis, P. A. Rolfe, K. T. Takusagawa, E. S. Lander, D. K. Gifford, E. Fraenkel, and R. A. Young. Transcriptional regulatory code of a eukaryotic genome. *Nature*, 431:99–104, Sep 2004.
- [Holland1977] P.W. Holland and R.E. Welsch. Robust regression using iteratively reweighted least-squares. *Communications in Statistics-Theory and Methods*, 6(9):813–827, 1977.
- [Hu2007] Z. Hu, P. J. Killion, and V. R. Iyer. Genetic reconstruction of a functional transcriptional regulatory network. *Nat. Genet.*, 39:683–687, May 2007.
- [Hu2009] Z. Hu, J. H. Hung, Y. Wang, Y. C. Chang, C. L. Huang, M. Huyck, and C. DeLisi. VisANT 3.5: multi-scale network visualization, analysis and inference based on the gene ontology. *Nucleic Acids Res.*, 37:W115–121, Jul 2009.
- [Huang2009a] d. a. W. Huang, B. T. Sherman, and R. A. Lempicki. Systematic and integrative analysis of large gene lists using DAVID bioinformatics resources. *Nat Protoc*, 4:44–57, 2009.
- [Huang2009b] d. a. W. Huang, B. T. Sherman, and R. A. Lempicki. Bioinformatics enrichment tools: paths toward the comprehensive functional analysis of large gene lists. *Nucleic Acids Res.*, 37:1–13, Jan 2009.
- [Huber2002] W. Huber, A. von Heydebreck, H. Sultmann, A. Poustka, and M. Vingron. Variance stabilization applied to microarray data calibration and to the quantification of differential expression. *Bioinformatics*, 18 Suppl 1:96–104, 2002.
- [Hughes2000] T. R. Hughes, M. J. Marton, A. R. Jones, C. J. Roberts, R. Stoughton, C. D. Armour, H. A. Bennett, E. Coffey, H. Dai, Y. D. He, M. J. Kidd, A. M. King, M. R. Meyer, D. Slade, P. Y. Lum, S. B. Stepaniants, D. D. Shoemaker, D. Gachotte, K. Chakraborty, J. Simon, M. Bard, and S. H. Friend. Functional discovery via a compendium of expression profiles. *Cell*, 102:109–126, Jul 2000.
- [Jeong2000] H. Jeong, B. Tombor, R. Albert, Z. N. Oltvai, and A. L. Barabasi. The large-scale organization of metabolic networks. *Nature*, 407:651–654, Oct 2000.
- [Jeong2001] H. Jeong, S. P. Mason, A. L. Barabasi, and Z. N. Oltvai. Lethality and centrality in protein networks. *Nature*, 411:41–42, May 2001.
- [Kanehisa2010] M. Kanehisa, S. Goto, M. Furumichi, M. Tanabe, and M. Hirakawa. KEGG for representation and analysis of molecular net-

- works involving diseases and drugs. *Nucleic Acids Res.*, 38:D355–360, Jan 2010.
- [Kaplan2008] N. Kaplan, I. K. Moore, Y. Fondufe-Mittendorf, A. J. Gossett, D. Tillo, Y. Field, E. M. LeProust, T. R. Hughes, J. D. Lieb, J. Widom, and E. Segal. The DNA-encoded nucleosome organization of a eukaryotic genome. *Nature*, 458:362–366, Mar 2009.
- [Khatri2007] P. Khatri, C. Voichita, K. Kattan, N. Ansari, A. Khatri, C. Georgescu, A. L. Tarca, and S. Draghici. Onto-Tools: new additions and improvements in 2006. *Nucleic Acids Res.*, 35:W206–211, Jul 2007.
- [Kitano2004] H. Kitano. Biological robustness. *Nat. Rev. Genet.*, 5:826–837, Nov 2004.
- [Krogan2006] N. J. Krogan, G. Cagney, H. Yu, G. Zhong, X. Guo, A. Ignatchenko, J. Li, S. Pu, N. Datta, A. P. Tikuisis, T. Punna, J. M. Peregrin-Alvarez, M. Shales, X. Zhang, M. Davey, M. D. Robinson, A. Paccanaro, J. E. Bray, A. Sheung, B. Beattie, D. P. Richards, V. Canadien, A. Lalev, F. Mena, P. Wong, A. Starostine, M. M. Canete, J. Vlasblom, S. Wu, C. Orsi, S. R. Collins, S. Chandran, R. Haw, J. J. Rilstone, K. Gandi, N. J. Thompson, G. Musso, P. St Onge, S. Ghanny, M. H. Lam, G. Butland, A. M. Altaf-Ul, S. Kanaya, A. Shilatifard, E. O’Shea, J. S. Weissman, C. J. Ingles, T. R. Hughes, J. Parkinson, M. Gerstein, S. J. Wodak, A. Emili, and J. F. Greenblatt. Global landscape of protein complexes in the yeast *Saccharomyces cerevisiae*. *Nature*, 440:637–643, Mar 2006.
- [Krushevskaya2009] D. Krushevskaya, H. Peterson, J. Reimand, M. Kull, and J. Vilo. VisHiC—hierarchical functional enrichment analysis of microarray data. *Nucleic Acids Res.*, 37:W587–592, Jul 2009.
- [Kull2008] M. Kull and J. Vilo. Fast approximate hierarchical clustering using similarity heuristics. *BioData Min*, 1:9, 2008.
- [Lee2003] S. I. Lee and S. Batzoglou. Application of independent component analysis to microarrays. *Genome Biol.*, 4:R76, 2003.
- [Lichtenberg2005] U. de Lichtenberg, L. J. Jensen, A. Fausb  ll, T. S. Jensen, P. Bork, and S. Brunak. Comparison of computational methods for the identification of cell cycle-regulated genes. *Bioinformatics*, 21:1164–1171, Apr 2005.
- [Liebermeister2002] W. Liebermeister. Linear modes of gene expression determined by independent component analysis. *Bioinformatics*, 18:51–60, Jan 2002.
- [MacIsaac2006] K. D. MacIsaac, T. Wang, D. B. Gordon, D. K. Gifford, G. D. Stormo, and E. Fraenkel. An improved map of conserved regulatory sites for *Saccharomyces cerevisiae*. *BMC Bioinformatics*, 7:113, 2006.

- [Margolin2006] A. A. Margolin, I. Nemenman, K. Basso, C. Wiggins, G. Stolovitzky, R. Dalla Favera, and A. Califano. ARACNE: an algorithm for the reconstruction of gene regulatory networks in a mammalian cellular context. *BMC Bioinformatics*, 7 Suppl 1:S7, 2006.
- [Markowetz2005] F. Markowetz, J. Bloch, and R. Spang. Non-transcriptional pathway features reconstructed from secondary effects of RNA interference. *Bioinformatics*, 21:4026–4032, Nov 2005.
- [Maslov2002] S. Maslov and K. Sneppen. Specificity and stability in topology of protein networks. *Science*, 296:910–913, May 2002.
- [Matys2006] V. Matys, O. V. Kel-Margoulis, E. Fricke, I. Liebich, S. Land, A. Barre-Dirrie, I. Reuter, D. Chekmenev, M. Krull, K. Hornischer, N. Voss, P. Stegmaier, B. Lewicki-Potapov, H. Saxel, A. E. Kel, and E. Wingender. TRANSFAC and its module TRANSCOMP: transcriptional gene regulation in eukaryotes. *Nucleic Acids Res.*, 34:D108–110, Jan 2006.
- [Metzker2010] M. L. Metzker. Sequencing technologies - the next generation. *Nat. Rev. Genet.*, 11:31–46, Jan 2010.
- [Minguez2009] P. Minguez, S. Gotz, D. Montaner, F. Al-Shahrour, and J. Dopazo. SNOW, a web-based tool for the statistical analysis of protein-protein interaction networks. *Nucleic Acids Res.*, 37:W109–114, Jul 2009.
- [Montaner2006] D. Montaner, J. Tarraga, J. Huerta-Cepas, J. Burguet, J. M. Vaquerizas, L. Conde, P. Minguez, J. Vera, S. Mukherjee, J. Valls, M. A. Pujana, E. Alloza, J. Herrero, F. Al-Shahrour, and J. Dopazo. Next station in microarray data analysis: GEPAS. *Nucleic Acids Res.*, 34:W486–491, Jul 2006.
- [Nadon2002] R. Nadon and J. Shoemaker. Statistical issues with microarrays: processing and analysis. *Trends Genet.*, 18:265–271, May 2002.
- [Oltvai2002] Z. N. Oltvai and A. L. Barabasi. Systems biology. Life’s complexity pyramid. *Science*, 298:763–764, Oct 2002.
- [Peer2001] D. Pe’er, A. Regev, G. Elidan, and N. Friedman. Inferring sub-networks from perturbed expression profiles. *Bioinformatics*, 17 Suppl 1:S215–224, 2001.
- [Przulj2004] N. Przulj, D. A. Wigle, and I. Jurisica. Functional topology in a network of protein interactions. *Bioinformatics*, 20:340–348, Feb 2004.
- [Radonjic2005] M. Radonjic, J. C. Andrau, P. Lijnzaad, P. Kemmeren, T. T. Kockelkorn, D. van Leenen, N. L. van Berkum, and F. C. Holstege. Genome-wide analyses reveal RNA polymerase II located upstream of genes poised for rapid response upon *S. cerevisiae* stationary phase exit. *Mol. Cell*, 18:171–183, Apr 2005.

- [Reimand2006] J. Reimand. Gene ontology mining tool GOST. Master's thesis, Institute of Computer Science, University of Tartu, Estonia, 2006.
- [Reimand2007] J. Reimand, M. Kull, H. Peterson, J. Hansen, and J. Vilo. g:Profiler—a web-based toolset for functional profiling of gene lists from large-scale experiments. *Nucleic Acids Res.*, 35:193–200, Jul 2007.
- [Reimand2008] J. Reimand, L. Tooming, H. Peterson, P. Adler, and J. Vilo. GraphWeb: mining heterogeneous biological networks for gene modules with functional significance. *Nucleic Acids Res.*, 36:W452–459, Jul 2008.
- [Reimand2010] J. Reimand, J. M. Vaquerizas, A. E. Todd, J. Vilo, and N. M. Luscombe. Comprehensive reanalysis of transcription factor knockout expression data in *Saccharomyces cerevisiae* reveals many new targets. *Nucleic Acids Res.*, in press, 2010.
- [Remenyi2004] A. Remenyi, H. R. Scholer, and M. Wilmanns. Combinatorial control of gene expression. *Nat. Struct. Mol. Biol.*, 11:812–815, Sep 2004.
- [Ren2000] B. Ren, F. Robert, J. J. Wyrick, O. Aparicio, E. G. Jennings, I. Simon, J. Zeitlinger, J. Schreiber, N. Hannett, E. Kanin, T. L. Volkert, C. J. Wilson, S. P. Bell, and R. A. Young. Genome-wide location and function of DNA binding proteins. *Science*, 290:2306–2309, Dec 2000.
- [Rung2002] J. Rung, T. Schlitt, A. Brazma, K. Freivalds, and J. Vilo. Building and analysing genome-wide gene disruption networks. *Bioinformatics*, 18 Suppl 2:S202–210, 2002.
- [Schena1995] M. Schena, D. Shalon, R. W. Davis, and P. O. Brown. Quantitative monitoring of gene expression patterns with a complementary DNA microarray. *Science*, 270:467–470, Oct 1995.
- [Schulz2009] H. Schulz, R. Kolde, P. Adler, I. Aksoy, K. Anastasiadis, M. Bader, N. Billon, H. Boeuf, P. Y. Bourillot, F. Buchholz, C. Dani, M. X. Doss, L. Forrester, M. Gitton, D. Henrique, J. Hescheler, H. Himmelbauer, N. Hubner, E. Karantzali, A. Kretsovali, S. Lubitz, L. Pradier, M. Rai, J. Reimand, A. Rolletschek, A. Sachinidis, P. Savatier, F. Stewart, M. P. Storm, M. Trouillas, J. Vilo, M. J. Welham, J. Winkler, A. M. Wobus, and A. K. Hatzopoulos. The FunGenES database: a genomics resource for mouse embryonic stem cell differentiation. *PLoS ONE*, 4:e6804, 2009.
- [Segal2003] E. Segal, M. Shapira, A. Regev, D. Pe'er, D. Botstein, D. Koller, and N. Friedman. Module networks: identifying regulatory modules and their condition-specific regulators from gene expression data. *Nat. Genet.*, 34:166–176, Jun 2003.

- [Simon2001] I. Simon, J. Barnett, N. Hannett, C. T. Harbison, N. J. Rinaldi, T. L. Volkert, J. J. Wyrick, J. Zeitlinger, D. K. Gifford, T. S. Jaakkola, and R. A. Young. Serial regulation of transcriptional regulators in the yeast cell cycle. *Cell*, 106:697–708, Sep 2001.
- [Smyth2004] G. K. Smyth. Linear models and empirical bayes methods for assessing differential expression in microarray experiments. *Stat Appl Genet Mol Biol*, 3:Article3, 2004.
- [Spellman1998] P. T. Spellman, G. Sherlock, M. Q. Zhang, V. R. Iyer, K. Anders, M. B. Eisen, P. O. Brown, D. Botstein, and B. Futcher. Comprehensive identification of cell cycle-regulated genes of the yeast *Saccharomyces cerevisiae* by microarray hybridization. *Mol. Biol. Cell*, 9:3273–3297, Dec 1998.
- [Svetlov1995] V. V. Svetlov and T. G. Cooper. Review: compilation and characteristics of dedicated transcription factors in *Saccharomyces cerevisiae*. *Yeast*, 11:1439–1484, Dec 1995.
- [Tavazoie1999] S. Tavazoie, J. D. Hughes, M. J. Campbell, R. J. Cho, and G. M. Church. Systematic determination of genetic network architecture. *Nat. Genet.*, 22:281–285, Jul 1999.
- [Tong2001] A. H. Tong, M. Evangelista, A. B. Parsons, H. Xu, G. D. Bader, N. Page, M. Robinson, S. Raghibizadeh, C. W. Hogue, H. Bussey, B. Andrews, M. Tyers, and C. Boone. Systematic genetic analysis with ordered arrays of yeast deletion mutants. *Science*, 294:2364–2368, Dec 2001.
- [Troyanskaya2001] O. Troyanskaya, M. Cantor, G. Sherlock, P. Brown, T. Hastie, R. Tibshirani, D. Botstein, and R. B. Altman. Missing value estimation methods for DNA microarrays. *Bioinformatics*, 17:520–525, Jun 2001.
- [Troyanskaya2005] O. G. Troyanskaya. Putting microarrays in a context: integrated analysis of diverse biological data. *Brief. Bioinformatics*, 6:34–43, Mar 2005.
- [Uetz2000] P. Uetz, L. Giot, G. Cagney, T. A. Mansfield, R. S. Judson, J. R. Knight, D. Lockshon, V. Narayan, M. Srinivasan, P. Pochart, A. Qureshi-Emili, Y. Li, B. Godwin, D. Conover, T. Kalbfleisch, G. Vijayadamodar, M. Yang, M. Johnston, S. Fields, and J. M. Rothberg. A comprehensive analysis of protein-protein interactions in *Saccharomyces cerevisiae*. *Nature*, 403:623–627, Feb 2000.
- [VanDriessche2005] N. Van Driessche, J. Demsar, E. O. Booth, P. Hill, P. Juván, B. Zupan, A. Kuspa, and G. Shaulsky. Epistasis analysis with global transcriptional phenotypes. *Nat. Genet.*, 37:471–477, May 2005.
- [Vaquerizas2004] J. M. Vaquerizas, J. Dopazo, and R. Diaz-Uriarte. DNMAID: web-based diagnosis and normalization for microarray data. *Bioin-*

- formatics*, 20:3656–3658, Dec 2004.
- [Vastrik2008] I. Vastrik, P. D’Eustachio, E. Schmidt, G. Joshi-Tope, G. Gopinath, D. Croft, B. de Bono, M. Gillespie, B. Jassal, S. Lewis, L. Matthews, G. Wu, E. Birney, and L. Stein. Reactome: a knowledge base of biologic pathways and processes. *Genome Biol.*, 8:R39, 2007.
- [Venables2002] W. N. Venables and B. D. Ripley. *Modern Applied Statistics with S*. Springer, New York, USA, 4th edition, 2003.
- [Vilo2002] J. Vilo. *Pattern Discovery from Biosequences*. PhD dissertation, Department of Computer Science, University of Helsinki, Finland, 2002.
- [Wasserman2004] W. W. Wasserman and A. Sandelin. Applied bioinformatics for the identification of regulatory elements. *Nat. Rev. Genet.*, 5:276–287, Apr 2004.
- [Yook2004] S. H. Yook, Z. N. Oltvai, and A. L. Barabasi. Functional and topological characterization of protein interaction networks. *Proteomics*, 4:928–942, Apr 2004.
- [Yu2008] H. Yu, P. Braun, M. A. Yildirim, I. Lemmens, K. Venkatesan, J. Sahalie, T. Hirozane-Kishikawa, F. Gebreab, N. Li, N. Simonis, T. Hao, J. F. Rual, A. Dricot, A. Vazquez, R. R. Murray, C. Simon, L. Tardivo, S. Tam, N. Svrikapa, C. Fan, A. S. de Smet, A. Motyl, M. E. Hudson, J. Park, X. Xin, M. E. Cusick, T. Moore, C. Boone, M. Snyder, F. P. Roth, A. L. Barabasi, J. Tavernier, D. E. Hill, and M. Vidal. High-quality binary protein interaction map of the yeast interactome network. *Science*, 322:104–110, Oct 2008.
- [Zhu2009] C. Zhu, K. J. Byers, R. P. McCord, Z. Shi, M. F. Berger, D. E. Newburger, K. Saulrieta, Z. Smith, M. V. Shah, M. Radhakrishnan, A. A. Philippakis, Y. Hu, F. De Masi, M. Pacek, A. Rolfs, T. Murthy, J. Labaer, and M. L. Bulyk. High-resolution DNA-binding specificity analysis of yeast transcription factors. *Genome Res.*, 19:556–566, Apr 2009.

ACKNOWLEDGEMENTS

First, I would like to thank my dear Liis for her valuable patience and support during numerous all-nighters, long travels and seven-day workweeks. I would like to express my gratitude to people who have mentored and motivated me during my journey towards the degree and introduced me to the fascinating field of computational biology, notably my long-time supervisor Prof. Jaak Vilo at the University of Tartu (UT), and Dr. Nicholas Luscombe and Dr. Juan Vaquerizas at the EMBL European Bioinformatics Institute (EBI). I am grateful for having worked with many great scientists at BIIT, EBI and elsewhere, who have provided stimulating environments, fruitful collaborations and co-authored papers. I would like to acknowledge Dr. Swen Laur for fair criticisms on an earlier manuscript of this dissertation, Anu Aun for conducting yeast experiments, and Hedi Peterson for sharing gradolence pain (the rat thing). Last but not least, my gratitude goes to my parents, family and friends.

This dissertation was financially supported by the Doctoral School of Information and Communication Technology (IKTDK), Centre of Excellence in Computer Science (EXCS), COBRED (LSHB-CT-2007-037730) and ENFIN (LSHG-CT-2005-518254) EU Framework 6 projects. I would like to acknowledge individual fellowships from the Marie Curie Biostar programme, the Tiger University programme of the Estonian Information Technology Foundation, and Ustus Agur and Artur Lind foundations, and the Erasmus programme. The CPU-intensive computations for this dissertation were carried out at the UT High Performance Computing Center as well as the EBI computing cluster.

SUMMARY IN ESTONIAN

Geenigruppide, võrgustike ja reguleerivate süsteemide funktsionaalne analüüs

Kaasaegne biotehnoloogia võimaldab kirjeldada elu aluseid suuremahulistes eksperimentides. Näiteks saab mõõta tuhandete geenide avaldumistasemeid, otsida DNA-st mustreid või kirjeldada interaktsioonivõrgustikke. Andmete interpreteerimine hõlmab arvutuslikke väljakutseid. Käesolev doktoritöö sisaldab meetodeid, arvutuslikke eksperimente ja tarkvara geenigruppide, võrgustike ja reguleerivate süsteemide analüüsiks. Funktsionaalne analüüs on tõlgenduslik ülesanne, mille käigus seostatakse eksperimentaalsed leiud faktidega bioloogilistest protsessidest, komponentidest, jne. Kuna teadmiste baas on ebatäielik, kasutatakse selleks statistilise rikastuse hinnangut. Töö käigus on arendatud kolm tööriista: g:Profiler, GraphWeb ja KEGGanim on veebipõhised vahendid geenigruppide, -võrgustike ja radade tõlgendamiseks. Samuti on logistilise regressiooni põhjal loodud algoritm eGLM, mis valib andmetest tugevaimaid ennustavaid efekte. Töö bioloogiline osa pühendub pagaripärmi *S. cerevisiae* regulatsiooni analüüsile. Esimalt koostati transkriptsioonifaktoreid (TF) kirjeldav andmestik ning valideeriti see funktsionaalse analüüsi teel. Seejärel rakendati andmestikul eGLM meetodit protsessispetsiifiliste TF-ide ennustamiseks. Arvutused hästikirjeldatud rakutsükli andmetele näitavad, et eGLM valib välja kõik põhiregulaatorid ning edestab standardseid meetodeid. Vähetuntud statsionaarfaasi (G_0) analüüsis leiti relevantseid funktsioone, s.h. vananemine ja metabolismi aeglustumine. Mõned ennustatud TF-id valideeruvad ka eksperimentaalselt: teatud pärmimutantidel on G_0 -s tervetest rakkudest oluliselt erinev elunevus.

Dissertatsiooni sissejuhatavas osas on lahti kirjutatud töö bioloogilised ja arvutuslikud alused. Bioloogilises sissejuhatuses on antud ülevaade transkriptsionaalse geeniekspressiooni põhimõtetest, ülegenoomsetest tehnoloogiatest, regulatsioonivõrgustikest ja funktsionaalse analüüsi bioloogilistest aspektidest. Arvutuslik sissejuhatatus sisaldab rikastuse testimiseks vajalikke statistilisi meetodeid, mitmese testimise põhimõtteid, ülevaadet regressioonimudelitest. Samuti on toodud töö käigus valminud eGLM algoritmi põhjalik kirjeldus ja vastavate simulatsioonide tulemused.

Dissertatsioonis sisalduvad üks avaldamata ja neli avaldatud artiklit.

1. Esimene artikkel "g:Profiler – a web-based toolset for functional profiling of gene lists from large-scale experiments" sisaldab geenilistide funktsionaalse analüüsi tarkvara g:Profiler kirjeldust ning infot relevantsete algo-

ritmide ja andmestike kohta.

2. Teise artikli "GraphWeb: mining heterogeneous biological networks for gene modules with functional significance" sisuks on geenivõrgustike topoloogilise ning funktsionaalse analüüsi ja integratsiooni meetodid ning vastava tarkvara GraphWeb kirjeldus.
3. Kolmas artikkel "KEGGanim: pathway animations for high-throughput data" kirjeldab bioloogiliste radade geeniekspressiooni visualiseerimise vahendit KEGGanim.
4. Neljandas artiklis "Comprehensive reanalysis of transcription factor knock-out expression data in *Saccharomyces cerevisiae* reveals many new targets" on kaasaegsete vahendite abil koostatud põhjalik mikrokiibi- ja funktsionaalne analüüs varasemalt avaldatud pärmi *S. cerevisiae* transkriptsiooni-faktorite kompendiumile ning näidatud uue andmestiku eeliseid.
5. Viiendas artiklis "Ensemble models of perturbation, nucleosome positioning and protein-DNA interactions reveal novel G0 regulators in *Saccharomyces cerevisiae*" (avaldamata) on loodud logistilisel regressioonil põhinev algoritm eGLM (Ensemble Generalised Linear Models), koostatud paljudest andmetest integreeritud pärmi geeniregulatsiooni andmestik ning ennustatud pärmi statsionaarfaasi geeniregulaatoreid.

PUBLICATIONS

CURRICULUM VITAE

General

- Name: Jüri Reimand
- Born: 08.10.1981, Tartu, Estonia, Estonian citizen
- Address: Tähe, 54-11, Tartu, Estonia
- Phone: +372 56636136 (GSM)
- Email: Juri.Reimand[@]ut.ee

Education

- 2006-10 – University of Tartu, Institute of Computer Science, PhD student
- 2005-06 – University of Tartu, Institute of Computer Science, *magister scientiarum* (MSc) in computer science
- 2004-05 – Johannes Kepler University (Linz, Austria), exchange student
- 2001-05 – University of Tartu, Institute of Computer Science, *baccalaureus scientiarum* (BSc) in computer science
- 1989-2000 – Tallinn English College, secondary education

Languages

- Estonian (native speaker), English (very good), German (moderate)

Work experience

- Since 2008 – University of Tartu, Institute of Computer Science, research fellow
- 2007-08 – EMBL European Bioinformatics Institute (EMBL-EBI, Hinxton Genome Campus, Cambridge, UK), Marie Curie Biostar fellow
- 2005-07 – University of Tartu, Institute of Computer Science, programmer
- 2004 – Aprate LTD, software developer
- 2003-04 – Inspirale.net LTD, software developer

Research and development

Research topics

- Bioinformatics, computational biology, gene regulation, machine learning

Publications

1. Reimand J, Vaquerizas JM, Todd AE, Vilo J, Luscombe NM: Comprehensive reanalysis of transcription factor knockout expression data in *Saccharomyces cerevisiae* reveals many new targets. *Nucleic Acids Res* (2010) Advance Access published online on April 12, 2010, doi:10.1093/nar/gkq232.
2. Adler P¹, Peterson H¹, Agius P, Reimand J, Vilo J: Ranking genes by their co-expression to subsets of pathway members. *Ann NY Acad Sci* (2009) 1158:1-13.
3. Krushevskaya D, Peterson H, Reimand J, Kull M, Vilo J: VisHiC – hierarchical functional enrichment analysis of microarray data. *Nucleic Acids Res* (2009), 37:W587-592.
4. Adler P¹, Kolde R¹, Kull M, Tkachenko A, Peterson H, Reimand J, Vilo J: Mining for coexpression across hundreds of datasets using novel rank aggregation and visualization methods. *Genome Biol* (2009), 10:R139.
5. Schulz H¹, Kolde R¹, Adler P, Aksoy I, Anastassiadis K, Bader M, Billon N, Boeuf H, Bourillot PY, Buchholz F, Dani C, Doss MX, Forrester L, Gitton M, Henrique D, Hescheler J, Himmelbauer H, Hubner N, Karantzali E, Kretsovali A, Lubitz S, Pradier L, Rai M, Reimand J, Rolletschek A, Sachinidis A, Savatier P, Stewart F, Storm MP, Trouillas M, Vilo J, Welham MJ, Winkler J, Wobus AM, Hatzopoulos AK: The FunGenES database: a genomics resource for mouse embryonic stem cell differentiation. *PLoS ONE* (2009), 4:e6804.
6. Reimand J¹, Tooming L¹, Peterson H, Adler P, Vilo J: GraphWeb: mining heterogeneous biological networks for gene modules with functional significance. *Nucleic Acids Res* (2008) 36:W452-459.
7. Adler P¹, Reimand J¹, Jänes J, Kolde R, Peterson H, Vilo J: KEGGanim: pathway animations for high-throughput data. *Bioinformatics* (2008) 24: 588-590.
8. 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* (2007) 35:W193-200.

Awards and fellowships

- 2009 – Ustus Agur Fellowship, Association of Information Technology and Telecommunications
- 2009 – Artur Lind Fellowship, Estonian Gene Centre
- 2007, 2008 – Marie Curie Biostar Fellowship, EMBL-EBI
- 2007 – PhD fellowship, Estonian Information Technology Foundation
- 2004, 2005 – Erasmus fellowship, Johannes Kepler University

¹Equal contribution

Üldandmed

- Ees- ja perekonnanimi: Jüri Reimand
- Sünniaeg ja -koht, kodakondsus: 08.10.1981, Tartu, Eesti, eestlane
- Aadress: Tähe, 54-11, Tartu, Eesti
- Telefon: +372 56636136 (GSM)
- Email: Juri.Reimand[@]ut.ee

Haridus

- 2006-10 – Tartu Ülikool, Arvutiteaduse instituut, doktorant
- 2005-06 – Tartu Ülikool, Arvutiteaduse instituut, *magister scientiarum* (MSc) informaatikas
- 2004-05 – Johannes Kepler University (Linz, Austria), vahetusüliõpilane
- 2001-05 – Tartu Ülikool, Arvutiteaduse instituut, *baccalaureus scientiarum* (BSc) informaatikas
- 1989-2000 – Tallinna Inglise Kolledž, keskharidus

Keelteoskus

- eesti keel (emakeel), inglise keel (väga hea), saksa keel (kesktase)

Töökogemus

- Alates 2008 – Tartu Ülikool, Arvutiteaduse instituut, teadur
- 2007-08 – EMBL Euroopa bioinformaatika instituut (EMBL-EBI, Hinxton Genome Campus, Cambridge, UK), Marie Curie Biostar teadur
- 2005-07 – Tartu Ülikool, Arvutiteaduse instituut, programmeerija
- 2004 – Arote LTD, tarkvaraarendaja
- 2003-04 – Inspiral.net LTD, tarkvaraarendaja

Teaduslik ja arendustegevus

Peamised uurimisvaldkonnad

- Bioinformaatika, arvutuslik bioloogia, geeniregulatsioon, masinõpe

Publikatsioonide loetelu

1. Reimand J, Vaquerizas JM, Todd AE, Vilo J, Luscombe NM: Comprehensive reanalysis of transcription factor knockout expression data in *Saccharomyces cerevisiae* reveals many new targets. *Nucleic Acids Res* (2010) avaldatud elektrooniliselt 12.04.2010, doi:10.1093/nar/gkq232.

2. Adler P¹, Peterson H¹, Agius P, Reimand J, Vilo J: Ranking genes by their co-expression to subsets of pathway members. *Ann NY Acad Sci* (2009) 1158:1-13.
3. Krushevskaya D, Peterson H, Reimand J, Kull M, Vilo J: VisHiC – hierarchical functional enrichment analysis of microarray data. *Nucleic Acids Res* (2009), 37:W587-592.
4. Adler P¹, Kolde R¹, Kull M, Tkachenko A, Peterson H, Reimand J, Vilo J: Mining for coexpression across hundreds of datasets using novel rank aggregation and visualization methods. *Genome Biol* (2009), 10:R139.
5. Schulz H¹, Kolde R¹, Adler P, Aksoy I, Anastassiadis K, Bader M, Billon N, Boeuf H, Bourillot PY, Buchholz F, Dani C, Doss MX, Forrester L, Gitton M, Henrique D, Hescheler J, Himmelbauer H, Hubner N, Karantzali E, Kretsovali A, Lubitz S, Pradier L, Rai M, Reimand J, Rolletschek A, Sachinidis A, Savatier P, Stewart F, Storm MP, Trouillas M, Vilo J, Welham MJ, Winkler J, Wobus AM, Hatzopoulos AK: The FunGenES database: a genomics resource for mouse embryonic stem cell differentiation. *PLoS ONE* (2009), 4:e6804.
6. Reimand J¹, Tooming L¹, Peterson H, Adler P, Vilo J: GraphWeb: mining heterogeneous biological networks for gene modules with functional significance. *Nucleic Acids Res* (2008) 36:W452-459.
7. Adler P¹, Reimand J¹, Jänes J, Kolde R, Peterson H, Vilo J: KEGGanim: pathway animations for high-throughput data. *Bioinformatics* (2008) 24: 588-590.
8. 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* (2007) 35:W193-200.

Saadud uurimistoetused ja stipendiumid

- 2009 – Eesti Info- ja Kommunikatsioonitehnoloogia Liidu Ustus Aguri nimeline stipendium
- 2009 – Geenikeskus SA Artur Linnu nimeline stipendium
- 2007, 2008 – Marie Curie Biostar stipendium, EMBL-EBI
- 2007 – EITSA Tiigriülikooli doktorandi stipendium
- 2004, 2005 – Erasmuse stipendium, Johannes Kepler Ülikool

¹Võrdne kontributsioon

DISSERTATIONES MATHEMATICAE UNIVERSITATIS TARTUENSIS

1. Mati Heinloo. The design of nonhomogeneous spherical vessels, cylindrical tubes and circular discs. Tartu 1991, 23p.
2. Boris Komrakov. Primitive actions and the Sophus Lie problem. Tartu 1991, 14p.
3. Jaak Heinloo. Phenomenological (continuum) theory of turbulence. Tartu 1992, 47p.
4. Ants Tauts. Infinite formulae in intuitionistic logic of higher order. Tartu 1992, 15p.
5. Tarmo Soomere. Kinetic theory of Rossby waves. Tartu 1992, 32p.
6. Jüri Majak. Optimization of plastic axisymmetric plates and shells in the case of Von Mises yield condition. Tartu 1992, 32p.
7. Ants Aasma. Matrix transformations of summability and absolute summability fields of matrix methods. Tartu 1993, 17p.
8. Helle Hein. Optimization of plastic axisymmetric plates and shells with piece-wise constant thickness. Tartu 1993, 29p.
9. Toomas Kiho. Study of optimality of iterated Lavrentiev method and its generalizations. Tartu 1994, 24p.
10. Arne Kokk. Joint spectral theory and extension of non-trivial multiplicative linear functionals. Tartu 1995, 166p.
11. Toomas Lepikult. Automated calculation of dynamically loaded rigidplastic structures. Tartu, 1995, 93p.(in Russian).
12. Sander Hannus. Parametrical optimization of the plastic cylindrical shells by taking into account geometrical and physical nonlinearities. Tartu, 1995, 74p (in Russian).
13. Sergei Tupailo. Hilbert's epsilon-symbol in predicative subsystems of analysis. Tartu 1996, 132p.
14. Enno Saks. Analysis and optimization of elastic-plastic shafts in torsion. Tartu 1996, 94p.
15. Valdis Laan. Pullbacks and flatness properties of acts. Tartu 1999, 88p.
16. Märt Pöldvere. Subspaces of Banach spaces having Phelps' uniqueness property. Tartu 1999, 72p.
17. Jelena Ausekle. Compactness of operators in Lorentz and Orlicz sequence spaces. Tartu 1999, 70p.
18. Krista Fischer. Structural mean models for analyzing the effect of compliance in clinical trials. Tartu 1999, 124p.
19. Helger Lipmaa. Secure and efficient time-stamping systems. Tartu 1999, 56p.

20. Jüri Lember. Consistency of empirical k -centres. Tartu 1999, 146p.
21. Ella Puman. Optimization of plastic conical shells. Tartu 2000, 102p.
22. Kaili Müürisep. Eesti keele arvutigrammatika: süntaks. Tartu 2000, 107p.
23. Varmo Vene. Categorical programming with inductive and conductive types. Tartu 2000, 116p.
24. Olga Sokratova. Ω -rings, their flat and projective acts with some applications. Tartu 2000, 121p.
25. Maria Zeltser. Investigation of double sequence spaces by soft and hard analytical methods. Tartu 2001, 153p.
26. Ernst Tungel. Optimization of plastic spherical shells. Tartu 2001, 90p.
27. Tiina Puolakainen. Eesti keele arvutigrammatika: morfoloogiline ühes-tamine. Tartu 2001, 137p.
28. Rainis Haller. $M(r, s)$ -inequalities. Tartu 2002, 77p.
29. Jan Villemson. Size-efficient interval time stamps. Tartu 2002, 81p.
30. Eno Tõnisson. Solving of expression manipulation exercises in computer algebra systems. Tartu 2002, 91p.
31. Mart Abel. Structure of Gelfand-Mazur algebras. Tartu 2003, 92p.
32. Vladimir Kuchmei. Affine completeness of some Ockham algebras. Tartu 2003, 99p.
33. Olga Dunajeva. Asymptotic matrix methods in statistical inference problems. Tartu 2003, 77p.
34. Mare Tarang. Stability of the spline collocation method for Volterra integro-differential equations. Tartu 2004, 88p.
35. Tatjana Nahtman. Permutation invariance and reparameterizations in linear models. Tartu 2004, 89p.
36. Märt Möls. Linear mixed models with equivalent predictors. Tartu 2004, 69p.
37. Kristiina Hakk. Approximation methods for weakly singular integral equations with discontinuous coefficients. Tartu 2004, 136p.
38. Meelis Käärik. Fitting Sets to Probability Distributions. Tartu 2005, 89p.
39. Inga Parts. Piecewise polynomial collocation methods for solving weakly singular integro-differential equations. Tartu 2005, 140p.
40. Natalia Saealle. Convergence and summability with speed of functional series. Tartu 2005, 90p.
41. Tanel Kaart. The reliability of linear mixed models in genetic studies. Tartu 2006, 123p.
42. Kadre Torn. Shear and bending response of inelastic structures to dynamic loads. Tartu 2006, 141p.
43. Kristel Mikkor. Uniform factorization for compact subsets of Banach spaces of operators. Tartu 2006, 72p.
44. Darja Saveljeva. Quadratic and cubic spline collocation for Volterra integral equations. Tartu 2006, 116p.
45. Kristo Heero. Path Planning and Learning Strategies for Mobile Robots in Dynamic Partially Unknown Environments. Tartu 2006, 122p.
46. Annely Mürk. Optimization of inelastic plates with cracks. Tartu 2006, 136p.

47. Annemai Raidjõe. Sequence spaces defined by modulus functions and superposition operators. Tartu 2006, 97p.
48. Olga Panova. Real Gelfand-Mazur algebras. Tartu 2006, 81p.
49. Härmel Nestra. Iteratively defined transfinite trace semantics and program slicing with respect to them. Tartu 2006, 118p.
50. Margus Pihlak. Approximation of multivariate distribution functions. Tartu 2007, 81p.
51. Ene Käärik. Handling dropouts in repeated measurements using copulas. Tartu 2007, 98p.
52. Artur Sepp. Affine models in mathematical finance: an analytical approach. Tartu 2007, 148p.
53. Marina Issakova. Solving of linear equations, linear inequalities and systems of linear equations in interactive learning environment. Tartu 2007, 169p.
54. Kaja Sõstra. Restriction estimator for domains. Tartu 2007, 103p.
55. Kaarel Kaljurand. Attempto Controlled English as a Semantic Web Language. Tartu 2008, 161p.
56. Mart Anton. Mechanical modeling of IPMC actuators at large deformations. Tartu 2008, 123p.
57. Evelyn Leetma. Solution of smoothing problems with obstacles. Tartu 2009, 81p.
58. Ants Kaasik. Estimating ruin probabilities in the Cramér-Lundberg model with heavy-tailed claims. Tartu 2009, 136p.
59. Reimo Palm. Numerical comparison of regularization algorithms for solving ill-posed problems. Tartu 2010, 105p.
60. Indrek Zolk. The commuting bounded approximation property of Banach spaces. Tartu 2010, 107 p.