Analysis of factors affecting the growth of oil bodies in A. Thaliana seeds: use of ordinary least squares and quantile regression

Ghassen Trigui, Martine Miquel, Bertrand B. Dubreucq, Olivier David, Alain Trubuil

Tampere International Center for Signal Processing. TICSP series # 63

Reija Autio, Ilya Shmulevich, Korbinian Strimmer, Carsten Wiuf, Septimia Sarbu &
Olli Yli-Harja (eds.)

The 10th International Workshop on Computational Systems Biology, WCSB 2013, June 10-12, Tampere, Finland

TAMPEREEN TEKNILLINEN YLIOPISTO
TAMPERE UNIVERSITY OF TECHNOLOGY

Tampere International Center for Signal Processing. TICSP series # 63

Reija Autio, Ilya Shmulevich, Korbinian Strimmer, Carsten Wiuf, Septimia Sarbu & Olli Yli-Harja (eds.)

The 10th International Workshop on Computational Systems Biology, WCSB 2013, June 10-12, Tampere, Finland

Tampere International Center for Signal Processing
Tampere 2013

ISBN 978-952-15-3091-3 (printed)
ISBN 978-952-15-3092-0 (PDF)
ISSN 1456-2774

The 10th International Workshop on Computational Systems Biology - WCSB 2013

During the last ten years, the Workshop on Computational Systems Biology (WCSB) has brought together the researchers across the world studying computational systems biology. The meeting has offered a multidisciplinary discussion forum for high quality multidisciplinary science, where many interdisciplinary bridges have been built. The WCSB is organized by the Computational Systems Biology Research Group in the Department of Signal Processing at Tampere University of Technology (TUT), and collaborating European research groups. During the years the meeting has expanded reflecting the rapid development in experimental biosciences and growth in the research of computational methods for systems biology.

The first four WCSB events in 2003 -2006 were organized in Tampere University of Technology, Finland. In 2008 the program committee set the target to render the event more international and since then WCSB has been organized together with our collaborators across Europe including

- WCSB 2008 hosted by the University of Leipzig, Germany,
- WCSB 2009 hosted by Aarhus University, Denmark,
- WCSB 2010 hosted by University of Luxembourg, Luxembourg,
- WCSB 2011 hosted by ETH Zürich, Switzerland, and
- WCSB 2012 hosted by Ulm University, Germany.

This year WCSB is celebrating its 10th year anniversary. The 10th International Workshop on Computational Systems Biology is brought back to its hometown Tampere, Finland. For the occasion, WCSB 2013 is proud to present 10 invited talks from internationally acknowledged experts of their respective fields in systems biology research. In addition, this year WCSB called for longer full length research papers with 3000 words, as well as abstracts. Further, this year WCSB is collaborating with the well-recognized publisher BioMed Central, and the best papers of the WCSB 2013 are published as Supplements to BMC Bioinformatics and BMC Systems Biology. Together 9 papers are accepted to be published in the Supplement to BMC Bioinformatics - Selected articles from the 10th International Workshop on Computational Systems Biology (WCSB) 2013: Bioinformatics, and 5 papers in Supplement to BMC Systems Biology - Selected articles from the 10th International Workshop on Computational Systems Biology (WCSB) 2013: Systems Biology.

This Proceedings of the 10th International Workshop on Computational Systems Biology - WCSB 2013, collects together the abstracts of the keynote and invited talks, the abstracts of the papers accepted to the BMC Supplements, the full research papers as well as the abstracts submitted to WCSB 2013.

We would like to thank the authors, the reviewers and the organizers for their contributions to WCSB and the proceedings. In addition, the financial support of The Federation of Finnish Learned Societies, Tampere Graduate School in Information Science and Engineering (TISE), Tampere International Center for Signal Processing (TICSP), and the Academy of Finland is gratefully acknowledged.

Reija Autio and Olli Yli-Harja

WCSB 2013 ORGANIZATION

Organizing institutions

Tampere International Center for Signal Processing(TICSP)
Tampere University of Technology, CSB - group
BioMediTech, Institute of Biosciences and Medical Technology

Chairs of WCSB

Ilya Shmulevich (Institute for Systems Biology)
Olli Yli-Harja (Tampere University of Technology)

BMC Supplement Editors

Ilya Shmulevich (Institute for Systems Biology)
Korbinian Strimmer(University of Leipzig)
Carsten Wiuf (University of Copenhagen)
Reija Autio (Tampere University of Technology)

Scientific Committee

Martin Bossert (Ulm University)
Frank Emmert-Streib (Queen's University Belfast)
Jose Fonseca (Uninova)
Heinz Koeppl (ETH Zurich)
Harri Lähdesmäki (Aalto University)
Ilya Shmulevich (Institute for Systems Biology)
Korbinian Strimmer (University of Leipzig)
Ralf Takors (University of Stuttgart)
Tapio Visakorpi (University of Tampere)
Carsten Wiuf (University of Copenhagen)
Wei Zhang (University of Texas, MD Anderson Cancer Center)
Olli Yli-Harja (Tampere University of Technology)
Miika Ahdesmäki (Almac Diagnostics)
Antti Honkela (University of Helsinki)
Alberto Sanz (University of Tampere)
Steffen Schober (Ulm University)
Reija Autio (Tampere University of Technology)
Juha Kesseli (Tampere University of Technology)
Matti Nykter (Tampere University of Technology)
Andre Ribeiro (Tampere University of Technology)
Meenakshisundaram Kandhavelu (Tampere University of Technology)

Local Organizers

Reija Autio (Tampere University of Technology)
Virve Larmila (Tampere University of Technology)
Eero Lihavainen (Tampere University of Technology)
Huy Tran (Tampere University of Technology)

Acknowledgements

We gratefully appreciate help from Ms. Virve Larmila, from Department of Signal Processing, Tampere University of Technology, for her devoted work in organization of the workshop. Special thanks are also due to Andre Ribeiro, Eero Lihavainen, Huy Tran, Maria Lehtivaara, Aansa Ali, Lingjia Kong and Matti Nykter for their diligent work in organization of the WCSB 2013. In addition, we thank the CSB group members from Department of Signal Processing, Tampere University of Technology, for their valuable help in organization of the workshop.

TABLE OF CONTENTS

KEYNOTE LECTURES

UNRAVELING PRINCIPLES OF GENE REGULATION USING THOUSANDS OF DESIGNED REGULATORY SEQUENCES

Eran Segal[1]

[1]Department of Computer Science And Applied Mathematics, Weizmann Institute of Science, Israel

eran.segal@weizmann.ac.il

ABSTRACT

Genetic variation in non-coding regulatory regions accounts for a significant fraction of changes in gene expression among individuals from the same species. However, without a 'regulatory code' that informs us how DNA sequences determine expression levels, we cannot predict which sequence changes will affect expression, by how much, and by what mechanism. To address this challenge, we developed a high-throughput method for constructing libraries of thousands of fully designed regulatory sequences and measuring their expression levels in parallel, within a single experiment, and with an accuracy similar to that obtained when each sequence is constructed and measured individually. Using this ~1000-fold increase in the scale with which we can study the effect of sequence on expression, we designed and measured the expression of libraries in which the location, number, affinity and organization of different types of regulatory elements has been systematically perturbed. Our results provide several new insights into principles of gene regulation, bringing us closer towards a mechanistic and quantitative understanding of which how expression levels are encoded in DNA sequence.

ASSEMBLY AND INTERROGATION OF TUMOR-SPECIFIC REGULATORY MODELS REVEALS MASTER REGULATORS OF TUMOR MAINTENANCE AND CHEMOSENSITIVITY

Andrea Califano

Department of Biomedical Informatics and Center for Computational Biology and Bioinformatics, Columbia University, New York, NY 10032, USA
califano@c2b2.columbia.edu

ABSTRACT

The recent onslaught of molecular data, across multiple human malignancies, is producing an unprecedented repertoire of genetic and epigenetic alterations contributing to tumorigenesis and progression. Yet, the direct impact of this knowledge on tumor treatment and prevention is still largely unproven. Loss of tumor suppressor function is difficult to target pharmacologically and, with a handful of exceptions, alterations providing potential drug targets are relatively infrequent in cancer patients and are thus unlikely to support clinical development.

By reconstructing and interrogating the in vivo regulatory logic of the cancer cell, which integrates multiple aberrant signals resulting from genetic and epigenetic alterations, systems biology is starting to elucidate and mechanistically validate both oncogene and non-oncogene addiction mechanisms. These mechanisms are exquisitely dependent on the molecular landscape of cancer subtypes, can be targeted pharmacologically, and are frequently synergistic, thus providing uniquely specific entry points for combination therapy.

In this presentation, we will discuss recent result in the discovery of synergistic, non-oncogene addiction mechanisms and their application to the stratification and treatment of high-grade glioma, non-small cell lung cancer, and prostate cancer. The approach is highly extensible and has been applied to a variety of additional tumor subtypes, to the study of stem cell differentiation, reprogramming, and pluripotency control, as well as to the study of neurodegenerative diseases.

INVITED TALKS

LEARNING SPATIAL AMINO ACID CONTACTS FROM MANY HOMOLOGOUS PROTEIN SEQUENCES

Erik Aurell[1]

[1]KTH Royal Institute of Technology, Stockholm, Sweden
eaurell@kth.se

ABSTRACT

Spatially proximate amino acids in a protein tend to co-evolve. A protein's three-dimensional (3D) structure hence leaves an echo of correlations in the evolutionary record. Reverse engineering 3D structures from such correlations is an open problem in structural biology, pursued with increasing vigor as more and more protein sequences continue to fill the data banks. Within this task lies a statistical inference problem, rooted in the following: correlation between two sites in a protein sequence can arise from firsthand interaction but can also be network-propagated via intermediate sites; observed correlation is not enough to guarantee proximity.

An approach to separating direct from indirect interactions is to learn a plausible probabilistic model from the data, and then score putative interactions by the corresponding terms in the model. In the context of protein sequences and learning the a model of at most pair-wise interactions (a Potts model) this approach has been referred to as direct-coupling analysis. The computational tasks involved are not trivial as in these problems a maximum likelihood approach is unfeasible, and one must resort to approximations.

I will discuss this field focusing on our recent result that the pseudolikelihood method significantly outperforms other existing approaches to the direct-coupling analysis.

This is joint work with Magnus Ekeberg, Cecilia Lövkvist, Yueheng Lan and Martin Weigt published in Phys. Rev. E 87, 012707 (2013), URL: http://link.aps.org/doi/10.1103/PhysRevE.87.012707 Code implementing the pseudolikelihood method for these problems is available at http://plmdca.csc.kth.se/.

GENOME-SCALE GENE REGULATORY NETWORKS IN BIOLOGY AND MEDICINE: FROM E. COLI TO BREAST CANCER

Frank Emmert-Streib[1]

[1] Computational Biology and Machine Learning Laboratory, Center for Cancer Research and Cell Biology, School of Medicine, Dentistry and Biomedical Sciences,
Queen's University Belfast, Belfast, UK.
f.emmert-streib@qub.ac.uk

ABSTRACT

Within recent years the availability of large-scale gene expression data enabled the inference of genome-scale regulatory networks for biological and biomedical data sets. Interestingly, the overlap between such networks and phenomenological networks like the protein interaction network and the transcriptional regulatory network has received little attention. For this reason, we provide an in-depth analysis of the structural, functional and chromosomal relationship between a protein interaction network, a transcriptional regulatory network and an inferred gene regulatory network, for S. cerevisiae and E. coli. As a result our study provides guidelines for the integration of different types of biological networks. Furthermore, we present regulatory networks for B cell lymphoma, breast cancer and colorectal cancer.

SEGMENTATION OF MICROCOPY IMAGES USING GRADIENT PATH LABELING AND ARTIFICIAL INTELLIGENCE TECHNIQUES

Jose Fonseca

Computational Intelligence Research Group, Uninova, Portugal
jmf@uninova.pt

ABSTRACT

Recent studies using microbes as model organisms rely on microscope imaging which needs to be complemented with reliable and fast methods of computer assisted image processing. These methods aim at facilitating the extraction of information from images of bacterial populations with single cell resolution, by avoiding manual analysis, which is fastidious, time consuming and subject to observer variances. To isolate single cells in microscopy images, image segmentation techniques are essential. However, segmentation of nontrivial images is one of the most difficult tasks in image processing. In this talk several segmentation methods will be compared and discussed. Artificial intelligence techniques to circumvent the over-segmentation typical of most classical segmentation methods will also be presented. Practical application examples will be shown to illustrate the results of the different techniques.

INTERDISCIPLINARY APPROACH TO TRACK RNA GRANULES MOVEMENT DURING CELL CYCLE IN YEAST

Cecilia Garmendia-Torres[1], Alexander Skupin[1,2], Sean Michael[1], Pekka Ruusuvuori[3], Nathan J. Kuwada[4], Didier Falconnet[5], Carl Hansen[5], Paul A. Wiggins[4], Aimée M. Dudley[1]

[1]Institute for Systems Biology, Seattle, USA

[2]Luxembourg Centre for Systems Biomedicine, University of Luxembourg, Esch-sur-Alzette, Luxembourg

[3]Department of Signal Processing, Tampere University of Technology, Tampere, Finland

[4]Physics and Bioengineering, University of Washington, Seattle, USA

[5]Centre for High-Throughput Biology, Department of Physics and Astronomy, University of British Columbia, Vancouver, Canada

To regulate translation, RNA transcripts can be stored in cytoplasmic granules that influence biological processes ranging from germ cell development to neuronal plasticity. For example, they transport maternal RNAs from the oocyte to the embryo and in mammalian neurons; neuronal granules transport RNA over long distances to the synaptic dendrites where translation can initiate in response to stimulation. The disruption of proper localization of these granules contributes to developmental and neurological diseases. A better knowledge of their subcellular movements is important to better understand how cells regulate post-transcriptional gene expression. That is the reason why we chose to study a subclass of RNA granules present in eukaryote cells called processing bodies (p-bodies).

To understand P-bodies involvement in RNA transport we measured the spatiotemporal dynamics of p-bodies over the course of the yeast cell cycle by integrating microfluidic-based single-cell imaging, automated image analysis, assessment of biophysical parameters, and genetics. Our results demonstrate that these granules undergo a unidirectional transport from the mother to the daughter cell during mitosis. This transport is dependent on the yeast RNA transport machinery composed of the myosin Myo4p, the protein adaptor She3p, and the RNA binding protein, She2p. We also find that p-bodies exhibit a constrained motion near the bud site as much as half an hour before the bud is observable and that this 'corralled' movement is also Myo4p and She2p dependent.

Our results imply that the mother cell sends RNAs to the daughter cell in a unidirectional manner under specific conditions and that these might participate in important aspects of cell division.

COMPUTATIONAL METHODS TO ANALYZE LARGE-SCALE AND HETEROGENEOUS DATA IN CANCERS

Sampsa Hautaniemi

Systems Biology Laboratory, Genome-Scale Biology Research Program,
University of Helsinki, Helsinki, Finland, P.O. Box 63 (Haartmaninkatu 8)
FI-00014 University of Helsinki
sampsa.hautaniemi@helsinki.fi

ABSTRACT

Systems level understanding of complex diseases requires coordinated efforts to collect and share genome-scale data from large patient cohorts. However, translating genome-scale data into knowledge and further to effective diagnosis, treatment and prevention strategies requires effective computational approaches that allow analysis and integration of multidimensional data with clinical parameters and knowledge available in bio-databases.

In this presentation I present an efficient computational ecosystem called Anduril that allows advanced analysis of large-scale biomedical data as well as integration to clinical data. I present examples on how Anduril based approach has resulted in biomedically interesting genes and their regulations in breast cancer, glioblastoma multiforme and prostate cancer.

SYSTEMS BIOLOGY APPROACHES TO STUDY TRANSCRIPTOMICS AND EPIGENETICS OF T CELL LINEAGE SPECIFICATION

Harri Lähdesmäki

Aalto University, Espoo, Finland
harri.lahdesmaki@aalto.fi

INTEGRATIVE SEQUENCING OF PROSTATE CANCER

Matti Nykter

Institute of Biomedical Technology, University of Tampere
matti.nykter@uta.fi

ABSTRACT

Prostate cancer is the most frequently diagnosed male cancer in developed nations. Most prostate tumors are slow growing and non-invasive, but some tumors develop an aggressive phenotype. Such tumors respond to androgen ablation, but eventually relapse with a castration resistant phenotype. The cause of this transformation is not fully understood, and no effective treatments exist. To shed light on the alterations giving rise to castration resistance we used integrative high throughput sequencing to study cancer-associated alterations in 53 prostate tumors at the DNA, RNA and epigenetic levels. The cohort included both untreated and castration resistant prostate cancers.

We identified a number of new genomic alterations and transcripts that are linked to prostate cancer progression, including two new functionally relevant fusion genes a number of novel prostate cancer associated transcripts, including transcripts specific to castration resistant tumors. Based on chip-seq data from prostate cancer cell lines, many of these novel transcripts are regulated by known oncogenes such as ERG and AR. Methylation sequencing revealed hypermethylation signature for castration resistant prostate cancer. Promoter hypermethylation suppressed the expression of hundreds of genes, but a subset of genes characterized by promoter H3K27 trimethylation responded to hypermethylation with increased expression.

TRANSCRIPTOME ANALYSIS USING NEXT-GENERATION SEQUENCING

Daniel Nicorici[1]

[1]Orion Pharma, Finland
daniel.nicorici@orionpharma.com

ABSTRACT

Rapid development of next-generation sequencing technologies provides an opportunity for systematic characterization of cell transcriptomes. Accessibility to transcriptome-scale information is energizing the study of RNA and it is having profound effects to fields such as molecular biology, medicine, biomedical research, bioinformatics, and evolutionary biology. High-throughput transcriptome sequencing (RNA-seq) has applications, such as gene/transcript/exon expression quantification at unprecendented scale, discovery of new events such as expressed fusion genes and alternative splicing, and indentification of expressed small scale mutations. RNA-seq has already deepened our understanding of gene fusions in cancer. An overview of different RNA-seq analysis methods is presented.

Supplement to BMC Bioinformatics - Selected articles from the 10th International Workshop on Computational Biology (WCSB) 2013: Bioinformatics

ZEBIAT, AN IMAGE ANALYSIS TOOL FOR REGISTERING ZEBRAFISH EMBRYOS AND QUANTIFYING CANCER METASTASIS

Teppo Annila[1], Eero Lihavainen[1*], Ines J. Marques[2], Darren R. Williams[3], Olli Yli-Harja[1] and Andre Ribeiro [1]*

[1] Department of Signal Processing, Tampere University of Technology, Tampere 33720, Finland,
[2] Cardiovascular Development and Repair, Centro Nacional de Investigacoes Cardiovasculares, Madrid, 28029, Spain,
[3] New Drug Targets Laboratory, School of Life Sciences, Gwangju Institute of Science and Technology, Gwangju, 500-712, Republic of Korea
*equal contribution
teppo.annila@tut.fi, andre.ribeiro@tut.fi

ABSTRACT

Zebrafish embryos have recently been established as a xenotransplantation model of the metastatic behaviour of primary human tumours. At the moment, the existing tools for automated data extraction from microscope images are restrictive concerning the developmental stage of the embryos, usually require laborious manual image preprocessing, and, in general, cannot characterize the metastasis as a function of the internal organs. We present a tool, ZebIAT, that allows both automatic or semi-automatic registration of the outer contour and inner organs of zebrafish embryos. The tool provides a registration at different stages of development and an automatic analysis of cancer metastasis per organ, thus allowing the study of cancer progression. The semi-automation relies on a graphical user interface. After validating the methods and exemplifying the usage of ZebIAT, we discuss its applicability. ZebIAT should be of use in high-throughput studies of cancer metastasis in zebrafish embryos.

CELL SEGMENTATION BY MULTI-RESOLUTION ANALYSIS AND MAXIMUM LIKELIHOOD ESTIMATION (MAMLE)

Sharif Chowdhury[1], Meenakshisundaram Kandhavelu[1], Olli Yli-Harja[1,2] and Andre S. Ribeiro[1]

[1] Department of Signal Processing, Tampere University of Technology, 33101 Tampere, Finland.
[2] Institute for Systems Biology, 401 Terry Avenue North, Seattle, WA 98109-5234, USA.
sharif.chowdhury@tut.fi, andre.ribeiro@tut.fi

ABSTRACT

Background: Cell imaging is becoming an indispensable tool for cell and molecular biology research. However, most of the studied processes are stochastic in nature, and require the observation of a large amount of cells and events. Ideally, extraction of information from these images ought to rely on automatic methods. Here, we propose a novel segmentation method, MAMLE, for detecting cells within dense clusters.

Methods: The proposed framework executes cell segmentation in two main stages. The first relies on state of the art filtering technique, edge detection in multi resolution with morphological operator and threshold decomposition for adaptive thresholding. From this initial segmentation result, a correction procedure is applied that exploits maximum likelihood estimate as an objective function. Also, it acquires morphological features from the initial segmentation for constructing the likelihood parameter, after which the final segmentation results are obtained.

Conclusions: We performed an empirical evaluation that includes sample images from different imaging modalities and diverse cell types. The new method attained very high (above 90%) cell segmentation accuracy in all test scenarios. Finally, its accuracy is compared to several recently proposed methods, and in all test samples, the method proposed here outperformed all of these in segmentation accuracy.

MULTI-SCALE GAUSSIAN REPRESENTATION AND OUTLINE-LEARNING BASED CELL IMAGE SEGMENTATION

Muhammad Farhan[1], Pekka Ruusuvuori[1], Mario Emmenlauer[2], Pauli Rämö[2], Christoph Dehio[2] and Olli Yli-Harja[1]

[1] Department of Signal Processing, Tampere University of Technology, 33720 Tampere, Finland
[2] Biozentrum, Universität Basel,4056 Basel, Switzerland
muhammad.farhan@tut.fi

ABSTRACT

Background: High-throughput genome-wide screening to study gene-specific functions, e.g. for drug discovery, demands fast automated image analysis methods to assist unlocking the full potential of such studies. Image segmentation is typically at the forefront of such analysis as performance of the subsequent steps, for example, cell classification, cell tracking etc., often relies on the results of segmentation.

Methods: We present a cell cytoplasm segmentation framework which first separates cell cytoplasm from image background using novel approach of image enhancement and coefficient of variation of multi-scale Gaussian scale-space representation. A novel outline learning-based classification method is developed using regularized logistic regression with embedded feature selection which classifies image pixels as outline/non-outline to give cytoplasm outlines. Refinement of the detected outlines to separate cells from each other is performed in a post-processing step where the nuclei segmentation is used as contextual information.

Results and conclusions: We evaluate the proposed segmentation methodology using two challenging test cases, presenting images with completely different characteristics, with cells of varying size, shape, texture and degrees of overlap. The feature selection and classification framework for outline detection produces very simple sparse models which use only a small subset of the large, generic feature set, that is, only 7 and 5 features for the two cases. Quantitative comparison of the results for the two test cases against state-of-the-art methods show that our methodology outperforms them with an increase of 4-9% in segmentation accuracy with maximum accuracy of 93%. Finally, the results obtained for diverse datasets demonstrate that our framework not only produces accurate segmentation but also generalizes well to different segmentation tasks.

A BAYESIAN APPROACH FOR PARAMETER ESTIMATION IN THE EXTENDED CLOCK GENE CIRCUIT OF ARABIDOPSIS THALIANA

Catherine F Higham[1] and Dirk Husmeier[1]

[1] School of Mathematics and Statistics, College of Science and Engineering, University of Glasgow, Glassgow G12 8QQ,Scotland, United Kingdom
Catherine.Higham@glasgow.ac.uk

ABSTRACT

The circadian clock is an important molecular mechanism that enables many organisms to anticipate and adapt to environmental change. Pokhilko et al. recently built a deterministic ODE mathematical model of the plant circadian clock in order to understand the behaviour, mechanisms and properties of the system. The model comprises 30 molecular species (genes, mRNAs and proteins) and over 100 parameters. The parameters have been fitted heuristically to available gene expression time series data and the calibrated model has been shown to reproduce the behaviour of the clock components. Ongoing work is extending the clock model to cover downstream effects, in particular metabolism, necessitating further parameter estimation and model selection. This work investigates the challenges facing a full Bayesian treatment of parameter estimation. Using an efficient adaptive MCMC proposed by Haario et al. and working in a high performance computing setting, we quantify the posterior distribution around the proposed parameter values and explore the basin of attraction. We investigate if Bayesian inference is feasible in this high dimensional setting and thoroughly assess convergence and mixing with different statistical diagnostics, to prevent apparent convergence in some domains masking poor mixing in others.

MODELING OF 2D DIFFUSION PROCESSES BASED ON MICROSCOPY DATA: PARAMETER ESTIMATION AND PRACTICAL IDENTIFIABILITY ANALYSIS

Sabrina Hock[1,2], Jan Hasenauer[1,2] and Fabian J Theis[1,2]

[1] Institute of Computational Biology, Helmholtz Zentrum München, Ingolstädter Landstr. 1, 85764 Neuherberg, Germany,
[2] Department of Mathematics, Technische Universität München, Boltzmannstr. 3, 85747 Garching, Germany
sabrina.hock@helmholtz-muenchen.de, fabian.theis@helmholtz-muenchen.de

ABSTRACT

Background: Diffusion is a key component of many biological processes such as chemotaxis, developmental differentiation and tissue morphogenesis. Since recently, the spatial gradients caused by diffusion can be assessed in-vitro and in-vivo using microscopy based imaging techniques. The resulting time-series of two dimensional, high-resolutions images in combination with mechanistic models enable the quantitative analysis of the underlying mechanisms. However, such a model-based analysis is still challenging due to measurement noise and sparse observations, which result in uncertainties of the model parameters.

Methods: We introduce a likelihood function for image-based measurements with log-normal distributed noise. Based upon this likelihood function we formulate the maximum likelihood estimation problem, which is solved using PDE-constrained optimization methods. To asses the uncertainty and practical identifiability of the parameters we introduce profile likelihoods for diffusion processes.

Results and Conclusion: As proof of concept, we model certain aspects of the guidance of dendritic cells towards lymphatic vessels, an example haptotaxis. Using a realistic set of artificial measurement data, we estimate the five kinetic parameters of this model and compute profile likelihoods. Our novel approach for the estimation of model parameters from image data as well as the proposed identifiability analysis approach is widely applicable to diffusion processes. The profile likelihood based method provides more rigorous uncertainty bounds in contrast to local approximation methods.

COMPUTING PREIMAGES OF BOOLEAN NETWORKS

Johannes Georg Klotz[1], Martin Bossert[1] and Steffen Schober[1]

[1] Institute of Communication Engineering, Ulm University, Albert-Einstein-Allee 43, 89081 Ulm, Germany
johannes.klotz@uni-ulm.de

ABSTRACT

In this paper we present an algorithm based on the sum-product algorithm that finds elements in the preimage of a feed-forward Boolean networks given an output of the network. Our probabilistic method runs in linear time with respect to the number of nodes in the network. We evaluate our algorithm for randomly constructed Boolean networks and a regulatory network of Escherichia coli and found that it gives a valid solution in most cases.

MAPPING BEHAVIORIAL SPECIFICATIONS TO MODEL PARAMETERS IN SYNTHETIC BIOLOGY

Heinz Koeppl[1,2], Marc Hafner[3] and James Lu[4]

[1] ETH Zurich, Zurich, Switzerland,
[2] IBM Zurich Research Laboratory, Rueschlikon, Switzerland,
[3] Harvard Medical School, Boston MA, USA,
[4] F. Hoffmann-La Roche, Basel, Switzerland
koeppl@ethz.ch

ABSTRACT

With the improvement of protocols for the assembly of transcriptional parts, synthetic biological devices can now be reliably assembled based on a design. The standardization of the parts open up the way for in silico design tools that improve the construct and optimize devices with respect to given formal specifications. The simplest such optimization is the selection of kinetic parameters and protein abundances such that the specified constraints are robustly satisfied. In this chapter we address the problem of determining parameter values that fulfill specifications expressed in terms of a functional on the trajectories of a dynamical model. We solve this inverse problem by linearizing the forward operator that maps parameter sets to specifications, and then inverting it locally. This approach has two advantages over brute-force random sampling. First, the linearization approach allows us to map back intervals instead of points and second, every obtained value in the parameter region is satisfying the specifications by construction.

EVALUATING A LINEAR K-MER MODEL FOR PROTEIN-DNA INTERACTIONS USING HIGH-THROUGHPUT SELEX DATA

Juhani Samuel Kähärä[1] and Harri Lähdesmäki[1,2]

[1] Department of Information and Computer Science, Aalto University School of Science, FI-00076 Aalto, Finland,
[2] Turku Centre for Biotechnology, University of Turku, Finland
juhani.kahara@aalto.fi

ABSTRACT

Transcription factor (TF) binding to DNA can be modeled in a number of different ways. It is highly debated which modeling methods are the best, how should the models be built and what can they be applied to. In this study a linear k-mer model proposed for predicting TF specificity in protein binding microarrays (PBM) is applied to a high-throughput SELEX data and the question of how to choose the most informative k-mers to the binding model is studied.

We implemented the standard cross-validation scheme to reduce the number of k-mers in the model and observed that the number of k-mers can often be reduced significantly without a great negative effect on prediction accuracy. We also found that the later SELEX enrichment cycles provide a much better discrimination between bound and unbound sequences as model prediction accuracies increased for all proteins together with the cycle number.

We compared prediction performance of k-mer and position specific weight matrix (PWM) models derived from the same SELEX data. Consistent with previous results on PBM data, performance of the k-mer model was on average 9 %-units better. For the 14 proteins in the SELEX data set classification accuracies were on average 71% and 62% for k-mer and PWMs, respectively.

Finally, the k-mer model trained with SELEX data was evaluated on ChIP-seq data demonstrating substantial improvements for some proteins. For GATA1 the model can distinguish between true ChIP-seq peaks and negative peaks. For RFX3 and NFATC1 data the model is as good as quessing.

CLASSIFICATION OF GENOMIC SIGNALS USING DYNAMIC TIME WARPING

Helena Skutkova[1], Martin Vitek[1,2], Petr Babula[2], Rene Kizek[3] and Ivo Provaznik[1,2]

[1] Department of Biomedical Engineering, Brno University of Technology, Technicka 12, CZ-61600 Brno, Czech Republic,
[2]International Clinical Research Center-Center of BIOmedical Engineering, St. Anne's University Hospital Brno, Pekarska 53, 65691 Brno, Czech Republic,
[3] Department of Chemistry and Biochemistry, Mendel University in Brno, Zemedelska 1, CZ-61300 Brno, Czech Republic
skutkova@feec.vutbr.cz, provaznik@feec.vutbr.cz

ABSTRACT

Background: Classification methods of DNA most commonly use comparison of the differences in DNA symbolic records, which requires the global multiple sequence alignment. This solution is often inappropriate, causing a number of imprecisions and requires additional user intervention for exact alignment of the similar segments. The similar segments in DNA represented as a signal are characterized by a similar shape of the curve. The DNA alignment in genomic signals may adjust whole sections not only individual symbols. The dynamic time warping (DTW) is suitable for this purpose and can replace the multiple alignment of symbolic sequences in applications, such as phylogenetic analysis.

Methods: The proposed method is composed of three main parts. The first part represent conversion of symbolic representation of DNA sequences in the form of a string of A,C,G,T symbols to signal representation in the form of cumulated phase of complex components defined for each symbol. Next part represents signals size adjustment realized by standard signal preprocessing methods: median filtration, detrendization and resampling. The final part necessary for genomic signals comparison is position and length alignment of genomic signals by dynamic time warping (DTW).

Results: The application of the DTW on set of genomic signals was evaluated in dendrogram construction using cluster analysis. The resulting tree was compared with a classical phylogenetic tree reconstructed using multiple alignment. The classification of genomic signals using the DTW is evolutionary closer to phylogeny of organisms. This method is more resistant to errors in the sequences and less dependent on the number of input sequences.

Conclusions: Classification of genomic signals using dynamic time warping is an adequate variant to phylogenetic analysis using the symbolic DNA sequences alignment; in addition, it is robust, quick and more precise technique.

Supplement to BMC Systems Biology - Selected articles from the 10th International Workshop on Computational Biology (WCSB) 2013: Systems Biology

BIOPROCESS DATA MINING USING REGULARIZED REGRESSION AND RANDOM FORESTS

Syeda Sakira Hassan[1], Muhammad Farhan[1*], Rahul Mangayil[2], Heikki Huttunen[1] and Tommi Aho[2]*

[1] Department of Signal Processing, Tampere University of Technology, Tampere, P.O.Box 553, 33101, Finland
[2] Department of Chemistry and Bioengineering, Tampere University of Technology, Tampere, P.O.Box 541, 33101, Finland
*equal contribution
sakira.hassan@tut.fi

ABSTRACT

Background: In bioprocess development, the needs of data analysis include (1) getting overview to existing data sets, (2) identifying primary control parameters, (3) determining a useful control direction, and (4) planning future experiments. In particular, the integration of multiple data sets causes that these needs cannot be properly addressed by regression models that assume linear input-output relationship or unimodality of the response function. Regularized regression and random forests, on the other hand, have several properties that may appear important in this context. They are capable, e.g., in handling small number of samples with respect to the number of variables, feature selection, and the visualization of response surfaces in order to present the prediction results in an illustrative way.

Results: In this work, the applicability of regularized regression (Lasso) and random forests (RF) in bioprocess data mining was examined, and their performance was benchmarked against multiple linear regression. As an example, we used data from a culture media optimization study for microbial hydrogen production. All the three methods were capable in providing a significant model when the five variables of the culture media optimization were linearly included in modeling. However, multiple linear regression failed when also the multiplications and squares of the variables were included in modeling. In this case, the modeling was still successful with Lasso (correlation between the observed and predicted yield was 0.69) and RF (0.91).

Conclusion: We found that both regularized regression and random forests were able to produce feasible models, and the latter was efficient in capturing the non-linearity in the data. In this kind of a data mining task of bioprocess data, both methods outperform multiple linear regression.

EFFECTS OF MULTIMERIZATION ON THE TEMPORAL VARIABILITY OF PROTEIN COMPLEX ABUNDANCE

Antti Häkkinen[1], Huy Tran[1], Olli Yli-Harja[1,2], Brian Ingalls[3] and Andre S. Ribeiro[1]

[1] Department of Signal Processing, Tampere University of Technology, P.O. box 553, 33101 Tampere, Finland,
[2]Institute for Systems Biology, 1441 North 34th Street, Seattle, Washington 98103-8904, USA,
[3] Department of Applied Mathematics, University of Waterloo,
200 University Avenue West, Waterloo, Ontario, Canada
antti.hakkinen@tut.fi, andre.ribeiro@tut.fi

ABSTRACT

We explore whether the process of multimerization can be used as a means to regulate noise in the abundance of functional protein complexes. Additionally, we analyze how this process affects the mean level of these functional units, response time of a gene, and temporal correlation between the numbers of expressed proteins and of the functional multimers. We show that, although multimerization increases noise by reducing the mean number of functional complexes it can reduce noise in comparison with a monomer, when abundance of the functional proteins are comparable. Alternatively, reduction in noise occurs if both monomeric and multimeric forms of the protein are functional. Moreover, we find that multimerization either increases the response time to external signals or decreases the correlation between number of functional complexes and protein production kinetics. Finally, we show that the results are in agreement with recent genome-wide assessments of cell-to-cell variability in protein numbers and of multimerization in essential and non-essential genes in Escherichia coli, and that the effects of multimerization are tangible at the level of genetic circuits.

IDENTIFICATION OF GENETIC MARKERS WITH SYNERGISTIC SURVIVAL EFFECT IN CANCER

Riku Louhimo[1], Marko Laakso[1], Tuomas Heikkinen[2], Susanna Laitinen[1], Pekka Manninen[3], Vladimir Rogojin[1], Minna Miettinen[1], Carl Blomqvist[4], Jianjun Liu[5], Heli Nevanlinna[2] and Sampsa Hautaniemi[1]

[1] Systems Biology Laboratory, Genome-Scale Biology Research Program, University of Helsinki, Helsinki, Finland, P.O.Box 63(Haartmaninkatu 8), FI-00014 University of Helsinki Finland

[2] Department of Obstetrics and Gynecology, Helsinki University Central Hospital, Helsinki, Finland, P.O.Box 700(Haartmaninkatu 8), FI-00020 HUS Finland

[3]CSC - IT Center for Science Ltd, Espoo, Finland, P.O.Box 405(Keilaranta 14), FI-02101 Espoo Finland

[4] Department of Oncology, Helsinki University Central Hospital, Helsinki, Finland, P.O.Box 180,FI-00020 HUS Finland

[5] Human Genetics, Genome Institute of Singapore, Singapore, 60 Biopolis Street 02-01 Singapore 138672

riku.louhimo@helsinki.fi, sampsa.hautaniemi@helsinki.fi

ABSTRACT

Cancers are complex diseases arising from accumulated genetic mutations that disrupt intracellular signaling networks. While several predisposing genetic mutations have been found, these individual mutations account only for a small fraction of cancer incidence and mortality. With large-scale measurement technologies, such as single nucleotide polymorphisms (SNP) microarrays, it is now possible to identify combinatorial effects that have significant impact on cancer patient survival. The identification of synergetic functioning SNPs on genome-scale is a computationally daunting task and requires advanced algorithms. We introduce a novel algorithm, Geninter, to identify SNPs that have synergetic effect on survival of cancer patients. Using a large breast cancer cohort we generate a simulator that allows assessing reliability and accuracy of Geninter and logrank test, which is a standard statistical method to integrate genetic and survival data. Our results show that Geninter outperforms the logrank test and is able to identify SNP-pairs with synergetic impact on survival.

A NEW MODEL TO SIMULATE AND ANALYZE PROLIFERATING CELL POPULATIONS IN BRDU LABELING EXPERIMENTS

Daniella Schittler[1], Frank Allgöwer[1] and Rob De Boer[2]

[1] Institute for Systems Theory and Automatic Control, University of Stuttgart, Pfaffenwaldring 9, 70569 Stuttgart, Germany,
[2] Department of Theoratical Biology, Utrecht University, Padualaan 8,3584 CH Utrecht, The Netherlands
daniella.schittler@ist.uni-stuttgart.de

ABSTRACT

Background: This paper presents a novel model for proliferating cell populations in labeling experiments. It is especially tailored to the technique of Bromodeoxyuridine (BrdU), which is taken up by dividing cells and thus accumulates with increasing division number during uplabeling. The study of the evolving label intensities of BrdU labeled cell populations is aimed at quantifying proliferation properties such as division and death rates.

Results: In contrast to existing models, our model considers a labeling efficacy that follows a distribution, rather than a uniform value. It thereby allows to account for noise as well as possibly space-dependent heterogeneity in the effective label uptake of the individual cells in a population. Furthermore, it enables more informative comparison with experimental data: The population-level label distribution is provided as a model output, thereby increasing the information content compared to existing models that give the fraction of labeled cells or the mean label intensity.

We employ our model to study some naturally arising examples of heterogeneity in label uptake, which are not covered by existing models. With simulations of noisy and spacially heterogeneous label uptake, we demonstrate that our model contributes a more realistic quantitative description of labeling experiments.

Conclusion: The presented model is to our knowledge the first one that predicts the full label distribution for BrdU labeling experiments. Thus, it can exploit more information, namely the full intensity distribution, from labeling measurements, and thereby opens up new quantitative insights into cell proliferation.

CHARACTERIZATION OF ABERRANT PATHWAYS ACROSS HUMAN CANCERS

Antti Ylipää[1], Olli Yli-Harja[1], Wei Zhang[2] and Matti Nykter[3]

[1] Department of Signal Processing, Tampere University of Technology, Korkeakoulunkatu 1, 33101 Tampere, Finland,
[2]The University of Texas MD Anderson Cancer Center, 1515 Holcombe Blvd, Houston, TX 77030,
[3] Institute for Biomedical Technology, University of Tampere, Biokatu 8, 33520 Tampere, Finland
antti.ylipaa@tut.fi

ABSTRACT

Background: Cancer is a broad group of genetic diseases which account for millions of deaths worldwide each year. Cancers are classified by various clinical, pathological and molecular methods, but even within a well-characterized disease, there is a significant inter-patient variability in survival, response to treatment, and other parameters. Especially in molecular level, tumours of the same category can appear significantly dissimilar due to complex combinations of genetic aberrations leading to a similar malignancy. We sought to extend the current classification methods by studying tumour heterogeneity at pathway level.

Methods: We computed the rate of alterations in 1994 pathways and 2210 tumours consisting of eight different cancers. Using gene set enrichment analysis, each sample was computed a pathway aberration profile that reflected its molecular state. The profiles were analysed together to infer the characteristic aberration rates for each pathway within each cancer. Subgroups of tumours defined by similar pathway aberrations were identified using clustering analyses. The pathway aberration and gene expression profiles of the subgroups were consecutively compared across all eight cancer types to search for similar tumours crossing the standard classification.

Results: We identified pathways and processes that were common to all cancers as well as traits that are unique to a cancer type or closely related cancers. Studying the gene expression patterns within the pathway context suggested potential alteration mechanisms. Clustering analysis revealed five clinically relevant subgroups of tumours in four cancers that exhibited significant differences in survival compared to others. The cross-cancer analysis of the subgroups resulted in the identification of tumours that shared potentially significant alterations.

Conclusions: This study represents the first effort to extend the molecular characterizations towards pathway level descriptions across the family of cancers. In addition to providing a proof-of-concept for single sample pathway aberration analysis in this context, we present a comprehensive pathway aberration dataset that can be used to study pathway aberration patterns within or across cancers. Significant similarities between subgroups of different cancers on pathway and gene expression levels provide interesting hypotheses for understanding variable drug response, or transferring treatments across diseases by identifying common druggable pathways or genes, for example.

FULL PAPERS

ASSESSMENT OF REGRESSION METHODS FOR INFERENCE OF REGULATORY NETWORKS INVOLVED IN CIRCADIAN REGULATION

Andrej Aderhold[1], *Dirk Husmeier*[2], *V. Anne Smith*[1], *Andrew J. Millar*[3], *and Marco Grzegorczyk*[4]

[1]School of Biology, University of St. Andrews, UK
[2] School of Mathematics and Statistics, University of Glasgow, UK
[3] SynthSys, University of Edinburgh, UK
[4] Department of Statistics, TU Dortmund University, Germany
Email: aa796@st-andrews.ac.uk, grzegorczyk@statistik.tu-dortmund.de

ABSTRACT

We assess the accuracy of three established regression methods for reconstructing gene and protein regulatory networks in the context of circadian regulation. Data are simulated from a recently published regulatory network of the circadian clock in *Arabidopsis thaliana*, in which protein and gene interactions are described by a Markov jump process based on Michaelis-Menten kinetics. We closely follow recent experimental protocols, including the entrainment of seedlings to different light-dark cycles and the knock-out of various key regulatory genes. Our study provides relative assessment scores for the comparison of state-of-the art regression methods, investigates the influence of systematically missing values related to unknown protein concentrations and mRNA transcription rates, and quantifies the dependence of the performance on the degree of recurrency.

1. INTRODUCTION

Plants have to carefully manage their resources. The process of photosynthesis allows them to utilize sunlight to produce essential carbohydrates during the day. However, the earth's rotation predictably removes sunlight, and hence the opportunity for photosynthesis, for a significant part of each day, and plants need to orchestrate the accumulation, utilisation and storage of photosynthetic products in the form of starch over the daily cycle to avoid periods of starvation, and thus optimise growth rates.

In the last few years, substantial progress has been made to model the central processes of circadian regulation, i.e. the mechanism of internal time-keeping that allows the plant to anticipate each new day, at the molecular level [1, 2]. Moreover, simple mechanistic models have been developed to describe the feedback between carbon metabolism and the circadian clock, by which the plant adjusts the rates of starch accumulation and consumption in response to changes in the light-dark cycle [3]. What is needed is the elucidation of the detailed structure of the molecular regulatory networks and signalling pathways of these processes, by utilization and integration of transcriptomic, proteomic and metabolic concentration profiles that become increasingly available from international research collaborations like Agrogenomics[1] and Timet[2].

The inference of molecular regulatory networks from postgenomic data has been a central topic in computational systems biology for over a decade. Following up on the seminal paper in [5], a variety of methods have been proposed [6], and several procedures have been pursued to objectively assess the network reconstruction accuracy [7, 8, 6]. The present study follows up on this work and extends it in four important respects. Firstly, to make the evaluation more targeted at the specific problem of inferring gene and protein interactions related to circadian regulation, we take the central circadian clock network in *Arabidopsis thaliana*, as published in [2] , as a ground truth for evaluation, and closely follow recent experimental protocols for data generation, including the entrainment of seedlings to different light-dark cycles, and the knock-out of various key regulatory genes. Secondly, to make the data generated from this network as realistic as possible, we model gene and protein interactions as a Markov jump process based on Michaelis-Menten kinetics. This is to be preferred over mechanistic models based on ordinary differential equations (used e.g. in [1]), as it captures the intrinsic stochasticity of molecular interactions. Thirdly, we assess the impact of missing values on the reconstruction task. Protein-gene interactions affect transcription rates, but both these rates as well as protein concentrations might not be available from the wetlab assays. In such situations, mRNA concentrations have to be taken as proxy for protein concentrations, and rates have to be approximated by finite difference quotients. For both approximations, we quantify the ensuing deterioration in network reconstruction accuracy. Fourthly and finally, we investigate the dependence of the network reconstruction accuracy on the degree of recurrency in the network. The central circadian clock network is densely connected with several tight feedback loops. However, we expect the regulatory network, via which the clock acts on carbon metabolism, to be sparser and with more feed-forward structures. In our study we therefore quan-

[1]https://agronomics.ethz.ch/
[2]http://timing-metabolism.eu/

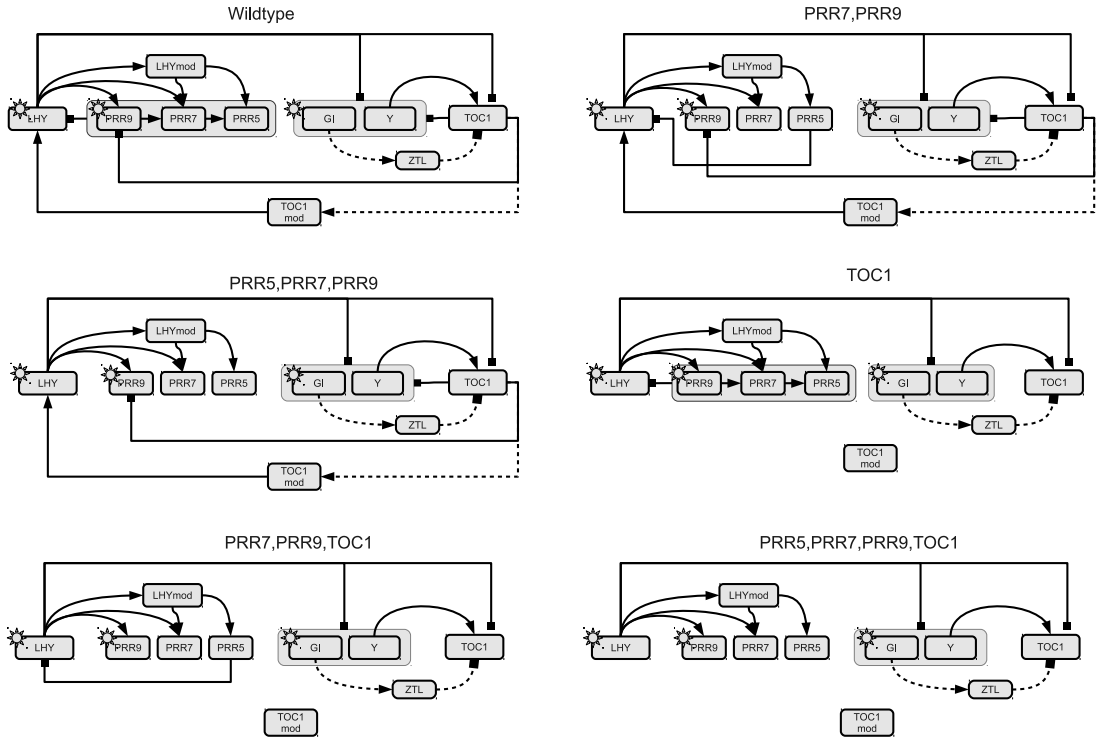


Figure 1. **Model network of the circadian clock in *Arabidopsis thaliana* based on [4] and subject to knock-outs.** Each graph shows interconnections of core circadian clock genes, where solid lines indicate protein influence on mRNA transcription, and dashed lines represent protein modifications. The top left panel shows the wildtye; the legends of the remaining panels indicate constant knock-outs of certain target proteins. Light influence is symbolized by a sun symbol, and grey boxes highlight set of regulators or regulated components.

tify how the network reconstruction depends on the degree of recurrency, and how the performance varies as critical feedback cycles are pruned.

2. METHOD OVERVIEW

2.1. Notation

Throughout the paper we use the following notation: For the regression models, which we will use to infer the network interactions, we have target variables y_g ($g = 1, \ldots, N$), each representing the temporal mRNA concentration gradient of a particular gene g. The realizations of each target variable y_g can then be written as a vector $\mathbf{y}_g = (y_{g,1}, \ldots, y_{g,T})^\mathsf{T}$, where $y_{g,t}$ is the realization of y_g in observation t. The potential covariates are either gene or protein concentrations, and the task is to infer a set of covariates $\boldsymbol{\pi}_g$ for each response variable y_g. The collective set of covariates $\{\boldsymbol{\pi}_1, \ldots, \boldsymbol{\pi}_N\}$ defines a regulatory interaction network, $\mathcal{M}$. In $\mathcal{M}$ the covariates and the target variables represent the nodes, and from each covariate in $\boldsymbol{\pi}_g$ a directed interaction (or "edge") is pointing to the target node g. The complete set of regulatory observations is contained in the design matrix $\mathbf{X}$. Realizations of the covariates in the set $\boldsymbol{\pi}_g$ are collected in $\mathbf{X}_{\pi_g}$, where the columns of $\mathbf{X}_{\pi_g}$ are the realizations of the covariates π_g. Design matrix $\mathbf{X}$ and $\mathbf{X}_{\pi_g}$ are extended by a constant element equal to 1 for the intercept.

2.2. Sparse regression

A widely applied linear regression method that encourages network sparsity is the Least Absolute Shrinkage and Selection Operator (Lasso) introduced in [9]. The Lasso optimizes the parameters of a linear model based on the residual sum of squares subject to an $L1$-norm penalty constraint on the regression parameters, $\|\mathbf{w}_g\|_1$, which excludes the intercept [10]:

$$\hat{\mathbf{w}}_g = \operatorname{argmin}\left\{ ||\mathbf{y}_g - \mathbf{X}^T \mathbf{w}_g||_2^2 + \lambda_1 \|\mathbf{w}_g\|_1 \right\} \quad (1)$$

where λ_1 is a regularisation parameter controlling the strength of shrinkage. Equation (1) constitutes a convex optimization problem, with a solution that tends to be sparse. Two disadvantages of the Lasso are arbitrary selection of single predictors from a group of highly correlation variables, and saturation at T predictor variables. To avoid these problems, the Elastic Net method was proposed in [11], which combines the Lasso penalty with a ridge regression penalty of the standard squared $L2$-norm $\|\mathbf{w}_g\|_2^2$ exluding the intercept:

$$\hat{\mathbf{w}}_g = \operatorname{argmin}\left\{ ||\mathbf{y}_g - \mathbf{X}^T \mathbf{w}_g||_2^2 + \lambda_1 \|\mathbf{w}_g\|_1 + \lambda_2 \|\mathbf{w}_g\|_2^2 \right\} \quad (2)$$

Like Equation (1), Equation (2) constitutes a convex optimization problem, which we solve with cyclical coordinate descent [10] implemented in the R software package *glmnet* . The regularization parameters λ_1 and λ_2 were optimized by 10-fold cross-validation.

2.3. Bayesian regression

In Bayesian regression we assume a linear regression model for the targets:

$$\mathbf{y}_g|(\mathbf{w}_g, \sigma_g, \pi_g) \sim \mathcal{N}(\mathbf{X}_{\pi_g}^{\mathsf{T}}\mathbf{w}_g, \sigma_g^2\mathbf{I}) \quad (3)$$

where σ_g^2 is the noise variance, and $\mathbf{w}_g$ is the vector of regression parameters, for which we impose a Gaussian prior:

$$\mathbf{w}_g|(\sigma_g, \delta_g, \pi_g) \sim \mathcal{N}(\mathbf{0}, \delta_g\sigma_g^2\mathbf{I}) \quad (4)$$

δ_g can be interpreted as a "signal-to-noise" hyperparameter [12]. For the posterior distribution we get:

$$\mathbf{w}_g|(\sigma_g, \delta_g, \pi_g, \mathbf{y}_g) \sim \mathcal{N}(\mathbf{\Sigma}_g\mathbf{X}_{\pi_g}\mathbf{y}_g, \sigma_g^2\mathbf{\Sigma}_g) \quad (5)$$

where $\mathbf{\Sigma}_g^{-1} = \delta_g^{-1}\mathbf{I} + \mathbf{X}_{\pi_g}\mathbf{X}_{\pi_g}^{\mathsf{T}}$, and the marginal likelihood can be obtained by application of standard results for Gaussian integrals [13]:

$$\mathbf{y}_g|(\sigma_g, \delta_g, \pi_g) \sim \mathcal{N}(\mathbf{0}, \sigma_g^2(\mathbf{I} + \delta_g\mathbf{X}_{\pi_g}^{\mathsf{T}}\mathbf{X}_{\pi_g})) \quad (6)$$

For σ_g^{-2} and δ_g^{-2} we impose conjugate gamma priors, $\sigma_g^{-2} \sim \mathrm{Gam}(\nu/2, \nu/2)$, and $\delta_g^{-1} \sim \mathrm{Gam}(\alpha_\delta, \beta_\delta)$.[3] The integral resulting from the marginalization over σ_g^{-2},

$$P(\mathbf{y}_g|\pi_g, \delta_g) = \int_0^\infty P(\mathbf{y}_g|\sigma_g, \delta_g, \pi_g)P(\sigma_g^{-2}|\nu)d\sigma_g^{-2}$$

is then a multivariate Student t-distribution with a closed-from solution (e.g. [13, 12]). Given the data for the potential covariates of y_g, symbolically $\mathcal{D}$, the objective is to infer the set of covariates π_g from the marginal posterior distribution:

$$P(\pi_g|\mathcal{D}, \mathbf{y}_g, \delta_g) = \frac{P(\pi_g)P(\mathbf{y}_g|\pi_g, \delta_g)}{\sum_{\pi_g^\star} P(\pi_g^\star)P(\mathbf{y}_g|\pi_g^\star, \delta_g)} \quad (7)$$

where the sum is over all valid covariate sets $\pi_g^\star$, $P(\pi_g)$ is a uniform distribution over all covariate sets subject to a maximal cardinality, $|\pi_g| \leq 3$, and δ_g is a nuisance parameter, which can be marginalized over. We sample sets of regulators (or covariates) π_g, signal-to-noise parameters δ_g, and noise variances σ_g^2 from the joint posterior distribution with Markov chain Monte Carlo (MCMC), following a Metropolis-Hastings within partially collapsed Gibbs scheme [12].

3. DATA

We generated data from the central circadian gene regulatory network in *Arabidopsis thaliana*, as proposed in [2] and depicted in the top right panel of Figure 1 . Following [14], the regulatory processes of transcriptional regulation and post-translational protein modification were described with a Markov jump process based on Michaelis-Menten kinetics, which defines how mRNA and protein concentrations change in dependence on the concentrations of other interacting components in the system (see appendix of [2] for detailed equations). We simulated mRNA and protein concentration time courses with the Gillespie algorithm [15], using the Bio-PEPA modelling framework [16]. To investigate the influence of recurrent interactions on the network reconstruction, we eliminated feedback loops successively via targeted downregulation of protein translation (knock-outs) and replacement of corresponding concentrations by white Gaussian noise. This gave us five modified network structures, as shown in Figure 1. For each network type we created 11 interventions in consistency with standard biological protocols (e.g. [17]). These include knock-outs of proteins 'GI', 'LHY', 'PRR7,PRR9', 'TOC1', and varying photoperiods of 4, 6, 8, 12, or 18 hours of light in a 24-hour light-dark (LD) cycle. For each intervention we simulated protein and mRNA concentration time courses over 6 days. The first 5 days served as entrainment to the indicated LD cycles. This was followed by a day of persistent darkness (DD) or light (LL), during which concentrations of mRNAs and proteins were measured in 2 hour intervals. Combining 13 observations for each intervention yielded 143 observations in total for each network type. All concentrations were standardized to unit standard deviation. The temporal mRNA concentration gradient was approximated by a difference quotient of mRNA concentrations based on two alternative temporal resolutions: at -2 and

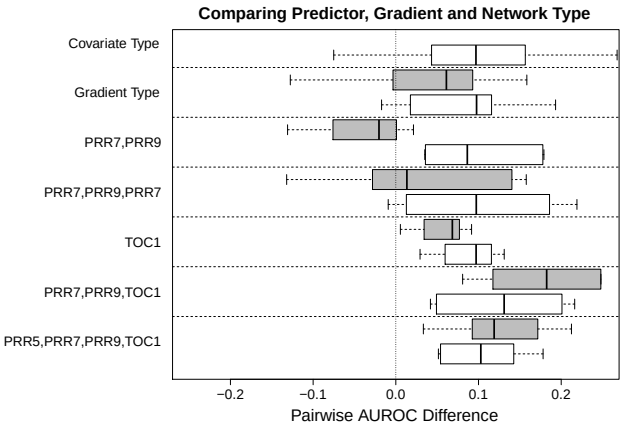


Figure 2. **Pairwise AUROC comparisons including all regression methods.** *Covariate type*: AUROC differences of protein against mRNA as covariates. *Gradient type*: AUROC difference of fine against coarse gradient for mRNAs (grey box) and proteins (white box). The remaining panels show the AUROC differences of various pruned networks as displayed in Figure 1 versus wild-type network for mRNA (grey boxes) and proteins (white boxes) as covariates.

[3] We set: $\nu = 0.01$, $\alpha_\delta = 2$, and $\beta_\delta = 0.2$, as in [12].

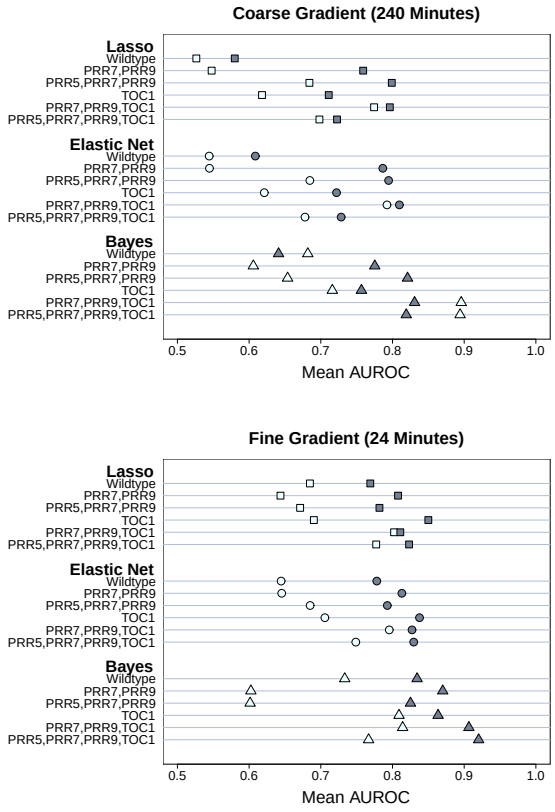


Figure 3. **Method comparison.** Mean AUROC values as a measure of network reconstruction performance for Lasso, Elastic Nets and Bayesian Regression applied to 6 distinct data types generated from networks shown in Figure 1. Empty symbols correspond to mRNA covariates, filled symbols to protein covariates.

+2 hours (coarse gradient), and at -12 and +12 minutes (fine gradient), followed by z-score standardization.

When trying to reconstruct the regulatory network from the simulated data, we ruled out self-loops, such as from LHY (modified) protein to LHY mRNA, and adjusted for mRNA degradation by enforcing mRNA self-loops, such as from the LHY mRNA back to itself. Protein 'ZTL' was included in the stochastic simulations, but excluded from structure learning because it has no direct effect on transcription. We carried out two different network reconstruction tasks. The first was based on complete observation, including both protein and mRNA concentration time series. The second was based on incomplete observation, where only mRNA concentrations were available, but protein concentrations were systematicaly missing. All network reconstructions were repeated on five independent data instantiations.

4. RESULTS

For Bayesian regression, we compute the marginal posterior probabilities of all potential interactions. For Lasso and Elastic Nets, we record the absolute values of non-zero regression parameters. Both measures provide a means by which interactions between genes and proteins

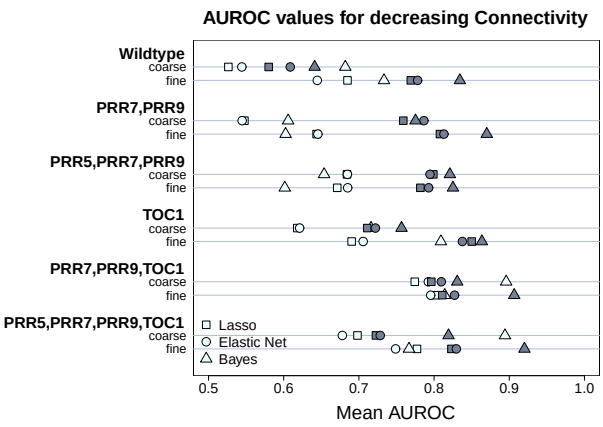


Figure 4. **Performance comparison in dependence on network connectivity.** Mean AUROC values for sparser networks in descending order for coarse (4 hour) and fine (24 minutes) response gradients. Empty symbols correspond to mRNA covariates, filled symbols to protein covariates.

can be ranked in terms of their significance or influence. Given that the true network is known, this ranking defines the Receiver Operating Characteristic (ROC) curve, where the sensitivity or recall is plotted against the complementary specificity. By numerical integration we then obtain the area under the curve (AUROC) as a global measure of network reconstruction accuracy, where larger values indicate a better performance, starting from AUROC=0.5 to indicate random expectation, to AUROC=1 for perfect network reconstruction. The results of our study are shown in Figures 2, 3 and 4 and can be summarized as follows.

Comparison between the methods. The performance of Lasso and Elastic Net is very similar, while Bayesian regression achieves slightly better results, especially when protein concentrations are included and temporal gradients are computed at fine resolution. This indicates a justification of the higher computational costs of inference based on MCMC.

Influence of gradient estimation. A finer temporal resolution for the gradient estimation tends to improve the network reconstruction. This affects in particular data for which the reconstruction based on coarse gradients is close to random expectation, and Bayesian regression models applied to protein concentrations. However, the coarse gradient leads to a noticeable improvement in the reconstruction with Bayesian regression based on mRNA profiles alone for the sparser networks in Figure 1. Preliminary investigations indicate that this unexpected trend is related to confounding correlations between the profiles of the two LHY isoforms, which are more important (propulsive) regulators in the sparser networks. However, a closer analysis is still required.

Influence of missing protein concentrations. With the noticeable exception of Bayesian regression applied to

data with coarse gradient estimation, discussed above, the inclusion of protein concentrations significantly improves the network reconstruction accuracy. Our study allows a quantification of the degree of improvement in terms of AUROC score differences, with a mean improvement of 0.09 and a p-value of 8e-06 (from a two a sided t-test) indicating significant higher values using protein covariates (Figure 2).

Influence of feedback loops. An important aspect of our study is the investigation of how the network reconstruction accuracy depends on the connectivity of the true network and the proportion of recurrent connections. To this end we have successively pruned feedback interactions, as shown in Figure 1. Figures 2 and 4 suggest that there is a noticeable trend, with less recurrent networks appearing to be easier to learn.

5. CONCLUSION

We have carried out a comparative evaluation of three established machine learning methods for regulatory network reconstruction (Lasso, Elastic Nets, Bayesian regression) based on the central gene regulatory network of the circadian clock in *Arabidopsis thaliana*, and a series of synthetic gene knock-outs that affect the proportion of recurrent interactions. Our study allows a quantification of the improvement in network reconstruction accuracy as a consequence of including protein concentrations, the dependence of the performance on the recurrent network connectivity, and the influence of the numerical approximation of the gradient (i.e. transcription rates) by finite-size difference quotients.

6. ACKNOWLEDGMENTS

This work was funded under EU FP7 project "Timet". M.G. is supported by the German Research Foundation (DFG), research grant GR3853/1-1. A.A. is supported by the BBSRC. SynthSys is a Centre for Integrative and Systems Biology partly funded by BBSRC and EPSRC award (BB/D019621).

7. REFERENCES

[1] A. Pokhilko, A. Fernández, K. Edwards, M. Southern, K. Halliday, and A. Millar, "The clock gene circuit in arabidopsis includes a repressilator with additional feedback loops," *Molecular systems biology* 8, 574, 2012.

[2] M. Guerriero, A. Pokhilko, A. Fernández, K. Halliday, A. Millar, and J. Hillston, "Stochastic properties of the plant circadian clock," *Journal of The Royal Society Interface* 9 (69), 744–756, 2012.

[3] F. Feugier and A. Satake, "Dynamical feedback between circadian clock and sucrose availability explains adaptive response of starch metabolism to various photoperiods," *Frontiers in Plant Science*, 3, 2012.

[4] A. Pokhilko et al., "Data assimilation constrains new connections and components in a complex, eukaryotic circadian clock model," *Molecular systems biology*, 6 (1), 2010.

[5] N. Friedman, M. Linial, I. Nachman, and D. Pe'er, "Using Bayesian networks to analyze expression data," *Journal of Computational Biology*, 7, 601–620, 2000.

[6] M. T. Weirauch et al., "Evaluation of methods for modeling transcription factor sequence specificity," *Nature Biotechnology*, 2013.

[7] D. Husmeier, "Sensitivity and specificity of inferring genetic regulatory interactions from microarray experiments with dynamic Bayesian networks," *Bioinformatics*, 19, 2271–2282, 2003.

[8] A. V. Werhli, M. Grzegorczyk, and D. Husmeier, "Comparative evaluation of reverse engineering gene regulatory networks with relevance networks, graphical Gaussian models and Bayesian networks," *Bioinformatics*, 22, 2523–2531, 2006.

[9] R. Tibshirani, "Regression shrinkage and selection via the lasso," *Journal of the Royal Statistical Society. Series B*, 58 (1), 267–288, 1995.

[10] J. Friedman, T. Hastie, and R. Tibshirani, "Regularization paths for generalized linear models via coordinate descent," *Journal of Statistical Software*, 33 (1), 1–22, 2010.

[11] H. Zou and T. Hastie, "Regularization and variable selection via the elastic net," *Journal of the Royal Statistical Society: Series B*, 67 (2), 301–320, 2005.

[12] M. Grzegorczyk and D. Husmeier, "A non-homogeneous dynamic Bayesian network with sequentially coupled interaction parameters for applications in systems and synthetic biology," *Statistical Applications in Genetics and Molecular Biology*, 11 (4), 2012a, Article 7.

[13] C. M. Bishop, *Pattern Recognition and Machine Learning*, Springer, Singapore, 2006.

[14] D. Wilkinson, *Stochastic modelling for systems biology*, 44, CRC Press, 2011.

[15] D. Gillespie, "Exact stochastic simulation of coupled chemical reactions," *The Journal of Physical Chemistry*, 81 (25), 2340–2361, 1977.

[16] F. Ciocchetta and J. Hillston, "Bio-pepa: A framework for the modelling and analysis of biological systems," *Theoretical Computer Science*, 410 (33), 3065–3084, 2009.

[17] K. Edwards et al., "Quantitative analysis of regulatory flexibility under changing environmental conditions," *Molecular Systems Biology*, 6 (1), 2010.

ORACLE CHARACTERIZATION FOR ACTIVE LEARNING FOR PROTEIN-PROTEIN INTERACTION PREDICTION

Seshan Ananthasubramanian[1,2], Jaime G. Carbonell Madhavi[3] and K. Ganapathiraju [1,2,3]

[1]Department of Biomedical Informatics and [2]Intelligent Systems program, University of Pittsburgh, 5607 Baum Blvd, Suite 401, Pittsburgh, PA, 15206, USA,

[3] Language Technologies Institute, Carnegie Mellon University, 5000 Forbes Ave, Pittsburgh, PA, 15213, USA,

madhavi@pitt.edu, madhavi@cs.cmu.edu

ABSTRACT

Discovery of human interactome is crucial for the understanding of complex biological processes and pathways, and for drug discovery. Bench-work experiments to determine protein-protein interactions (PPIs) are costly in terms of time, materials, equipment, and technical and scientific expertise. Thus, in recent years computational machine learning methods have been proposed to predict PPIs. These methods require training data to learn the classification models; but, currently known interactions are not sufficient to create an accurate model to predict the whole interactome or even a significant fraction thereof. If it is possible to seek additional training data experimentally, which are the protein-pairs that should be chosen to yield maximally-informative instances so as to maximize prediction accuracy for a given cost/effort level? Active machine learning (AL) methods are designed to select instances (protein-pairs) to create optimal training data and thus maximize their value in terms of prediction accuracy. Improving the relative accuracy is the pursuit of research in AL. However AL assumes the existence of an omniscient oracle (an experiment or an expert), which would give the correct label for every instance that it is asked, and it would do so for both positive and negative labels. In reality in the context of PPIs, an "oracle" can provide labels for only some interacting pairs (reluctance), and cannot give labels for non-interacting pairs (one-class nature). We develop algorithms for AL appropriate to this scenario of PPI prediction. Our results are superior compared to both a random baseline and also generic state-of-the-art AL.

1. INTRODUCTION

Computationally discovering protein-protein interactions (PPIs) of the human interactome is a challenging task for many reasons. Supervised machine learning algorithms have been applied to PPI prediction, which treat the task as a binary classification problem. Positive class data, namely the interacting protein-pairs, are only one-in-a-thousand or less from amongst 500 million possible pairs [1]. There are no known negative-class data, namely virtually no probably non-interacting pairs. Therefore, a classification model must be built using training data that consists of a pool of known interactions and a pool of randomly paired proteins not known to interact that are treated as non-interacting pairs [2]. The best estimates of the total number of interactions [3] indicate that only 5-10% are currently known, and it is possible that these interactions are not representative of the entire interaction space – i.e. they are not drawn randomly or i.i.d. from the interacting distribution. For more than half of all the proteins, there is not even one known interaction. In order to learn an accurate PPI classifier, one must sample the space and determine labels of those instances by experimental methods. However, acquiring more interactions would require performing bench work experiments, including medium to high-throughput methods such as yeast 2-hybrid [4] or mass spectrometry techniques [5], which are costly and time consuming processes that also require considerable amount of resources, high-end equipment and technical expertise. Therefore, designing strategies that optimally select instance pairs that are most informative for incrementally training an accurate classifier is an important goal in PPI prediction. This is known as active learning.

Active learning strategies help to select optimal training instances to achieve superior predictive accuracy within a given budget that is available for labeling instances [6]. See Figure 1. In active learning, the algorithm starts with the few labeled instances that are available ('🔴' and '🔵' in Figure 1). Next, it identifies the unlabeled instances (pairs of proteins) whose labels when known would prove most useful in learning a better classifier ('⊗' in Figure 1); an oracle is queried to obtain labels of those instances; the labels thus obtained are added to the training data and the model is re-trained to arrive at a more accurate classifier; if time and budget permit, this process is iterated. The basic assumption is that obtaining labels involves investment of resources, and therefore that the selected instances whose labels are asked should be optimal for retraining more accurate classifier. In PPI prediction, an oracle would typically be a bench-work experiment, which can characterize a protein-pair. Common strategies for active learning include density-based selection where more representative instances are selected from denser clusters [7], or uncertainty-based selection where data points are selected from maximum

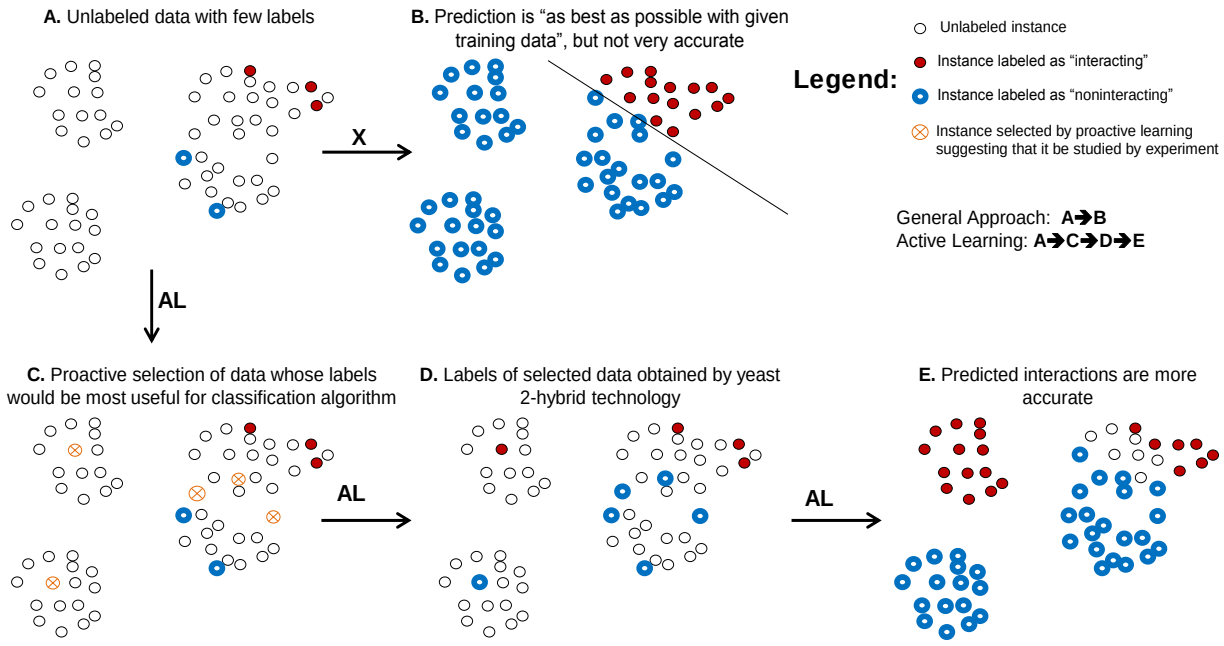


Figure 1 – Active learning concept diagram: Contrasting active learning (A-C-D-E) against normal supervised learning (A-B). When a few data points are selected and their labels are asked from an oracle (orange '+' instances), the classifier learnt after adding those instances to the training data is more accurate than without acquiring those labels. A major focus of research in active learning is on how to select the '+' instances so that maximum accuracy may be obtained under budget restrictions.

confusion or uncertain regions of the instance space with respect to the current classifier [8], or ensemble-methods which employ multiple criteria to select data points that typically outperform many other strategies [9]. "Wrapper" methods select those instances which would lead to the highest improvement in classifier accuracy once they are added to the training dataset [6] by the computationally-intensive step of hypothesizing the label of each instance, retraining the classifier as if this hypothesis were true, and estimating accuracy gains (for PPI this would require a half-billion retraining steps to select a single instance for experimentation; hence it is not a tractable option). We previously applied a variety of active learning methods including density based, uncertainty based and history based active selection approaches for predicting PPIs [10] in which we observed that active learning required only 500 labeled instances to achieve the same or superior accuracy as achieved by 3,000 randomly selected labeled instances. A six-fold improvement in experimental efficiency with modest computational effort is always a desirable tradeoff, but can we do better? This paper argues the positive.

2. APPROACH

Active learning assumes that there exists a single perfect oracle, which would *always* give the *correct answer* for labeling instances – e.g. an experimental procedure that always yields an answer and is always correct – this is unrealistic of experiments to determine protein-protein interactions. These characteristics of the oracle present a hindrance to optimal active learning for interactome-scale discovery of PPIs. Would active learning work better with a proper "oracle" for PPI? This is the central hypothesis of this paper, which we answer in the positive. The primary contributions of this work are (a) a detailed characterization of the oracle in the domain of PPI prediction, and (b) development of suitable active learning approaches to suit these oracle characteristics, and an empirical demonstration of their effectiveness over the state of the art.

2.1. Characteristics of the PPI oracle

The PPI oracle is not able to detect non-interacting proteins (one-class nature): Traditional bench-work experiments cannot validate "non-interaction nature" of protein pairs. There is no data available which gives us information about experimentally confirmed non-interacting proteins. Thus, PPI oracles can only provide labels for protein-pairs associated with a single class, which is the interacting class

The PPI oracle is only able to label a subset of all real interactions (reluctance): There are various technological constraints which limit the experiment to detect all possible interactions. Some experiments cannot be used on a certain class of proteins. For example, a yeast 2-hybrid (Y2H) experiment cannot be used for detecting interactions of a protein that is able to initiate transcription without its interacting partner or those that

are dependent on post-translational modifications; mass-spectrometry methods might fail to discover transient interactions; most experimental techniques that are available today are not able to characterize interactions involving integral membrane proteins [11]. Thus, technical constraints constitute a major roadblock for many experimental methods, and they are only able to identify a subset of all possible interactions.

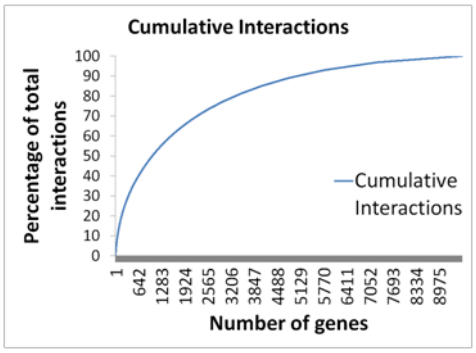


Figure 2 – Number of known interactions of human genes: Graph showing percentage of known interactions against the total number of genes associated with those interactions, sorted in descending order of the number of interactions. 50% known interactions are associated with 4% of genes.

The PPI oracle labels a subset of interactions incorrectly (fallibility)*:* Human errors and operation issues while performing the experiment can lead to incorrect labeling of some protein-pairs. A two hybrid assay can produce some biologically irrelevant interactions, especially if the proteins reside in different tissues or different sub cellular locations [11]. It is also difficult to isolate the binary interactions between protein-pairs when a protein complex is involved, as it is highly difficult to identify the target protein in the complex. Many high-throughput interactions are detected *in-vivo* by causing disruption of normal cellular function [11]. Thus non-typical interactions may be observed, as the existing pattern of protein interactions is disrupted resulting in the generation of false positives.

2.2. Active learning with PPI oracles

We extend general active learning methods to the scenario where the oracle has *one class* and *reluctance* properties, considering these characteristics one at a time. We consider the oracle to be *infallible* in nature; that is, if it does give a label, it is believed to be the correct label. We call this Active Learning with *Reluctant One Class* oracle. Whereas the reluctance property and the accuracy of oracles were introduced and characterized by Donmez et al [12], the one-class property is a new contribution to active learning driven by the requirements of PPI prediction.

2.2.1. Active learning with one class oracle

In classical active learning, if an oracle is resultant (no answer), we assume the majority class, i.e. the protein pair is non-interacting.. However in one class active learning, if the oracle fails to give the label for a data point, then we *estimate* whether a potential label can be associated with it by observing the oracle behavior. We assign the estimated label to the data point and add it to the training data instead of always assuming it to be non-interacting in nature. We also estimate *P(label | x)*, which is the probability that the data-point *x* would be labeled by the oracle. As the oracle provides the label for *only* the interacting class, we can use *P (label | x)* as a reasonable estimate to assign the "interacting" label to the instance. The unlabeled instance is assigned the interacting label with a probability *P (label | x)*.

There exists no real world datasets from which we can learn the behavior of the PPI oracle to calculate *P(label | x)* for unlabeled instances. Hence, we propose three different heuristic methods which provide a reasonable estimate of *P(label | x)* based on the distribution of known labeled data that have already been added to the training data in the active learning process. These heuristic methods are described in the next section. The following are assumptions in determining *P(label | x)* through heuristic based methods:

a. We assume that if the oracle gives the (interacting) label for a particular data-point, then it is highly likely to assign the same label for neighboring points in PP feature space. As the distance from known labeled instances increases, the likelihood of an instance being labeled by the oracle decreases.

b. Some proteins have been studied extensively due to their significant role in a significant pathway, or their role as a drug target or due to disease-association. For 9673 genes out of 22,500 genes there is at least one known interaction (Human Protein Reference Database: www.HPRD.org [13]); 60% of all genes have *no known* experimentally determined interaction. Figure 2 shows number of interactions of proteins ordered by their rank when ordered descending by the number of interactions known. It can be seen from the figure that around 50% of all known interactions are associated with only 1,250 genes or 4% of all genes. Thus *most* known interactions today are associated with a *very few* proteins. The labeled subset space is generally associated with those 4% of genes that are well studied in literature.

2.3. Active learning with reluctant oracle

When an oracle is asked for labels of unlabeled instances (*U*), of which N are interacting pairs in reality, the oracle gives labels for only *n* out of N interacting pairs owing to its reluctant nature described earlier. For example, in a yeast 2-hybrid set up, the "assay sensitivity" was found to be only 23% [14]. In order that all unlabeled interactions be recovered by the learning algorithm, it is expected that these *n* labels be a random subset from the N interactions for which the label was asked. If the

labeled positive examples are indeed a random subset of all positive examples for which labels are asked, then the true conditional probability that an instance is interacting, given by *P(y = "interacting"|x)* and probability that the instance is labeled, given by *P(label | x)*, differ only by a constant factor [15]. Thus,

$$P(y = "interacting"|x) = c * P(label|x) \quad \ldots (1)$$

If the subset of instances labeled by the oracle is representative of the interaction feature space, then many unknown interactions can be identified by the heuristic methods.

3. METHODS

3.1. Data

We used the dataset developed by Qi. et. al [2]. This dataset consists of 14,608 pairs of proteins that were known to interact. It also consists of 432,197 non-overlapping unlabeled instances. The protein-pairs (data instances) are represented by 27 features computed from biophysical characteristics of individual proteins. The 27 features correspond to: Gene Ontology (GO) cellular component, molecular function and biological process (3 features), co-occurrence in tissues (1 feature), gene co-expression (16 features), sequence similarity (1 feature), homology based (5 features) and domain interaction (1 feature). GO features measure the number of GO terms that are common in the annotations of the two proteins in the pair. As GO terms are categorized into three categories, namely the cellular component, molecular function and the biological process, the protein-pair features are computed separately for these three different similarity values. The tissue feature is a binary value indicating whether the two proteins have been expressed in the same tissue or not. This feature is added as it is observed that interacting proteins are likely to be expressed in the same tissue. 16 gene expression features were computed as correlations between gene expression values of the two genes in 16 different experiments. Sequence similarity is computed using BlastP sequence alignment E-value for the two proteins in the pair. Homologous proteins are obtained for each protein-pair in four different organisms namely yeast, fly, mouse and worm. This feature value is set to one, if the corresponding homologs are found to interact with each other in one or more of these organisms. Further details about the features in this dataset are described in the original source by Qi et. al [2].

This dataset consisted of 14,608 interactions and 432,197 random pairs. From the random pairs, we created a subset of instances from this data such that every pair has more than 50% feature coverage. This was done so as to maintain a balance of feature coverage between interacting and random pairs, as the interacting pairs had a better feature coverage than the random pairs [10]. This subset is used for the development and the evaluation of the proposed methods. This subset has 180,800 protein-pairs in total. All the known protein interactions in the original set were included in this subset. 160,800 protein-pairs were selected randomly from this dataset for training and another 20,000 for testing. The test-data contained 5% interactions, which constitute 1,000 interacting protein-pairs, while the training data contained the remaining 13,608 interactions. A skewed dataset distribution is used to mimic the realistic scenario.

3.2. Base Classifier

Previously, Bayesian classifiers, logistic regression, support vector machines, decision trees and random forest have been proposed as supervised learning classifiers for PPI prediction [2, 16, 17]. It has been shown that random forest is best suited for this domain [2]. We used a Random forest containing 20 trees built by choosing from 8 different random features.

3.3. Oracle Simulation

We simulate the oracle behavior for PPI predictions using the set of known interactions that was downloaded from the human protein reference database (HPRD). The simulated oracle would assign the "interacting" label to a data-point if the protein-pair associated with the point is listed in HPRD. HPRD lists about 38,000 interactions pooled from various experimental sources. This list forms about 5% of all possible human PPIs [3]. Thus HPRD can be thought of as a reluctant one class oracle which gives the labels for only 5% of all the interacting class.

3.4. Accuracy Metrics

Precision, recall and F-score *of positive class* will be plotted as a function of the total cost in active learning. Precision is measured as the fraction of correctly predicted protein interactions among all the pairs predicted by the classifier to be interacting. Recall is the fraction of the interacting protein-pairs which the classifier is able to correctly identify as interacting pairs. F-score is the harmonic mean of precision and recall. F-score measures the accuracy of the method by combining both precision and recall values. Hence it can be used as a measure to compare the accuracy of the methods.

3.5. Algorithms

We propose *active learning with a one class reluctant oracle*, which attempts to learn the behavior of the oracle, and uses it to estimate the positive labels that are associated with unlabeled points by estimating *P(label |x)*. We present three different ways to *estimate* the missing class for selected data points, which can be used in conjunction with any underlying active learning method. In the following methods, clusters are created using k-means clustering with Euclidean distance.

A. Estimate P(label | x) using number of interactions uncovered from each cluster ('cluster-interactions')

In this method, *P(label | x)* is estimated based on the number of labeled instances obtained from each cluster during the active learning iterations. The protein-pairs are clustered based on the existing feature-space, by considering all the features. If an oracle is able to provide the labels for many points in a cluster, it is most likely to do so for the other points too. That is, if the oracle which has a 25% recall value, is able to give the labels for even 15% of all protein-pairs in a cluster, then in reality 60% (i.e. 15*1/0.25) other unlabeled points in the same cluster are more likely to be interacting points. Based on this principle, we propose a new metric to estimate labels of unlabeled points in a cluster, based on the size of the cluster, the recall value of the oracle, and the total number of positive instances that have been uncovered from the cluster till the current iteration. This value increases with increase in the number of labeled instances from every cluster. *P(label | x)* is thus estimated adaptively, during every active learning iteration as follows:

$$P(label\ |x) = (1/Z) * P_k / L_k \ldots (2)$$

where, Z is a parameter whose value is between 0 and 1 and is chosen proportional to the recall of the oracle. As interacting class is a rare category in protein-protein interactions, if the oracle is highly reluctant, then setting value of Z to larger values would prove to be beneficial. k is the cluster to which point x belongs, P_k is the total number of labels obtained from the reluctant oracle, that is, the number of interactions that are uncovered from the cluster k. L_k is the total number of data-points belonging to the cluster k, whose labels are asked for by the system.

B. Estimate P(label | x) using distance from known uncovered interactions ('distance-interactions')

This method assigns P(label | x) based on the distance of the unlabeled data-point, from known uncovered interactions.

$$P(label\ |x) = Z * (d_m - \|x_i - x\|) / d_m \ldots (3)$$

where x_i is the nearest labeled interaction from the data-point x that has been added to the training instances during the course of the active-learning iterations, $\|x_i - x\|$ is the Euclidean distance between the points x_i and x; and d_m stands for the maximum distance between any data-point x and its closest interaction instance x_i.

Using this approach, the closer the points are to known interacting data-points, the more likely is the chance that they would be labeled as interactions. As distance from known interacting protein-pairs increases, this probability value reduces significantly. Z is a constant value between 0 and 1 which is assigned based on the recall value associated with the oracle.

C. Estimate P(label | x) using ε-neighborhood of uncovered interactions ('ε-neighborhood')

This method is based on the assumption that points closer to interacting data-points are more likely to be interacting in nature. We assign a label of "interacting" to all those unlabeled data-points shortlisted by any traditional active learning algorithm which falls in the ε-neighborhood of known uncovered interactions. The ε-neighborhood NN(x) of an instance consists of all those points located at most at a distance of ε from the data-point.

$$NN(x) = \{y | y \in U,\ \|x - y\| \leq \varepsilon\} \ldots (4)$$

Intuitively it could be thought of as a set of all data-points encompassed by a sphere with radius ε, drawn from the considered instance. For other points that do not fall in the ε-neighborhood of known interactions, we assign them to be "non-interacting" in nature.

Creation of a Weighted-Dataset

We also experimented with creation of a weighted-dataset. Instead of assigning the "interacting" label with a probability of *P(label |x)* to the unlabeled data-point, and then adding it to the training set, we can create a weighted-dataset, by weighing unlabeled examples with a probability *P(label |x)*. That is, each unlabeled example is considered to be "interacting" with a weight *P(label | x)* and "non-interacting" with a weight *1-P(label | x)*. Labeled data points are considered to be "known" interactions with a unit weight.

During the Active learning process, labels for select data-points are requested from the oracle. For all those points for which the label is obtained, a unit weight is assigned to each of them and they are added back to the training set. The unlabeled examples are duplicated. One copy is made as "interacting" with weight *P(label |x)* and the other is made as "non-interacting" with weight 1-*P(label | x)*. Both these copies are added to the training set and a classifier is trained on the same. This entire process is repeated in every iteration.

We identify two separate methods namely, *distance-interactions-weighted* and *distance-cluster-weighted* which estimate *P(label | x)* using the distance-interactions and the distance-cluster based methods respectively, but create a weighted-dataset instead of directly estimating the missing class.

4. RESULTS

As a first step in evaluating the one class active learning methods, we determined the best possible values that could be associated with Z and $\boldsymbol{\varepsilon}$ for all the three heuristics using a tenfold cross-validation technique on the training data. The values of Z and $\boldsymbol{\varepsilon}$ were chosen in such a manner that those values contributed to the highest increase in F-score, while maintaining a high value of precision above a chosen threshold of 60%. The details of the cross-validations carried out to select Z and $\boldsymbol{\varepsilon}$ (for each of the heuristic methods separately) are given in Supplementary File 1.

4.1. Evaluation on Test Data

After finding suitable values of Z and ε for each of the three heuristic methods separately, we evaluated the proposed one class heuristics on the held out data set and compared the performance against traditional active learning baseline and with random selection method. Uncertainty based active learning algorithm was used as the baseline, and 100 points were selected in every iteration. In the first iteration, 50 points are selected randomly and an initial classifier is built. Data is clustered into 50 clusters for the cluster-interactions method. Precision, recall and F-scores are computed for every iteration as shown in Figure 3.

It is observed that all of the active learners achieve the same F-score with only 1,500 instances, as the random selector does with 18,000 instances. Also the one class active learning methods have a higher value of F-score as compared to the baseline of uncertainty based active learning. For the parameters selected through cross-validation for each method, cluster-interactions method has the highest value of F-score, followed by the distance-interactions method and the ε-neighborhood method. The ε-neighborhood method has the same F-score as that of the baseline active learner for about 10,000 instances, after which the F-score associated with the baseline approach starts decreasing slightly. This decrease can be attributed to a biased selection of positive instances during the initial iterations of the active learning process. In the later iterations, more number of random unlabeled instances are picked for which the oracle is not able to provide the label and are added to the training set as negative data, which results in a small decrease in the F-score. However, the one class methods do not consider these points to be negative in nature. An appropriate label is assigned to such points using the heuristics proposed above. Some of the unlabeled data get labeled as positive instances due to which the F-score remains the same for these one class methods.

The challenge is to significantly improve upon recall of the entire system without compromising on its precision. The one class methods have a slightly lower value of precision but have much better values of recall, which can be observed through their increased F-scores. In the domain of PPI prediction, where the positive class is a very rare category, an improvement in recall is more difficult to attain than improvement in precision (because a slight inaccuracy in classifier could potentially misclassify an order of magnitude more of negative instances into positive class, thereby dropping precision significantly). Note that the reported precision and recall are computed for the positive class as is typical in this domain.

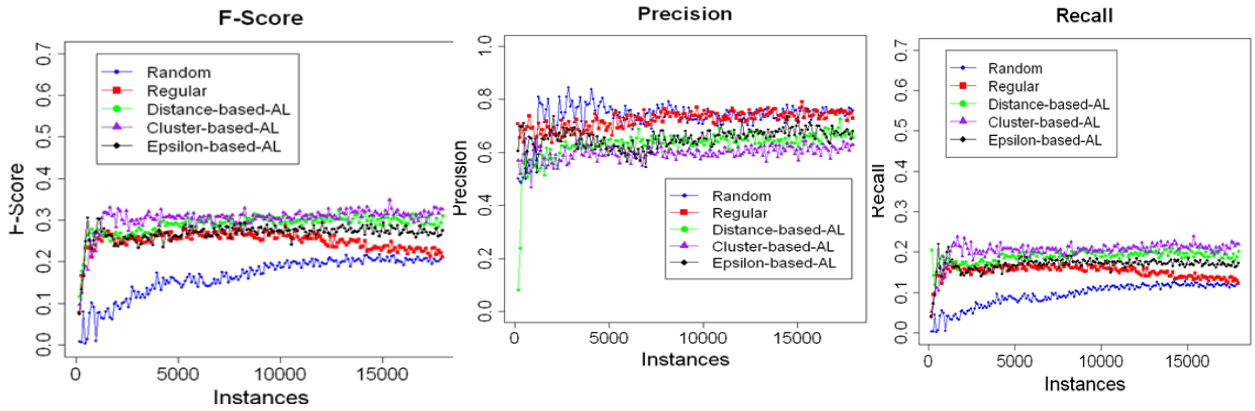


Figure 3 – Results of one class active learning by different approaches: Plots show results of comparison of one-class active learning methods with the active learning baseline and random-selection of instances using F-Score, precision and recall values of the positive class. It is the characteristic of this domain that the recall values are very low on account of positive class being a very rare category.

Figure 4 shows the results associated with the weighted-dataset methods. These methods are compared against the traditional active learning baseline as well as their corresponding counterparts that estimate potential labels before adding the data-point to the training set. Weighted-dataset methods are similar to the other one class active learning methods discussed previously. The *interactions-distance weighted* method has the same precision, recall and F-score trends as its one class active learning counterparts. The *interactions-cluster weighted* method has a very poor value of precision, and a very high value of recall as compared to its one class counterpart for the initial few iterations, and bounces back to , thus having a similar trend in the F-scores.

5. DISCUSSION

In the domain of protein-protein interaction prediction, or any computational biology task in general, the few instances of labeled data currently known are often insufficient to confidently characterize remaining unlabeled data. Conversely, data characterization using wet-lab experiments is expensive in terms of expert

manpower, time, and resources. When choosing a specific data instance (e.g. molecule or protein-pair) to be studied by wet-lab experiments, its redundancy with previously annotated (labeled) data is rarely taken into account. The choice of which molecule is to be studied is mostly based on the expertise of the lab proposing to study it and on availability of reagents that make it possible to study with the said experimental methods. However, the trend today is moving towards high-throughput techniques (HT) such as yeast 2-hybrid

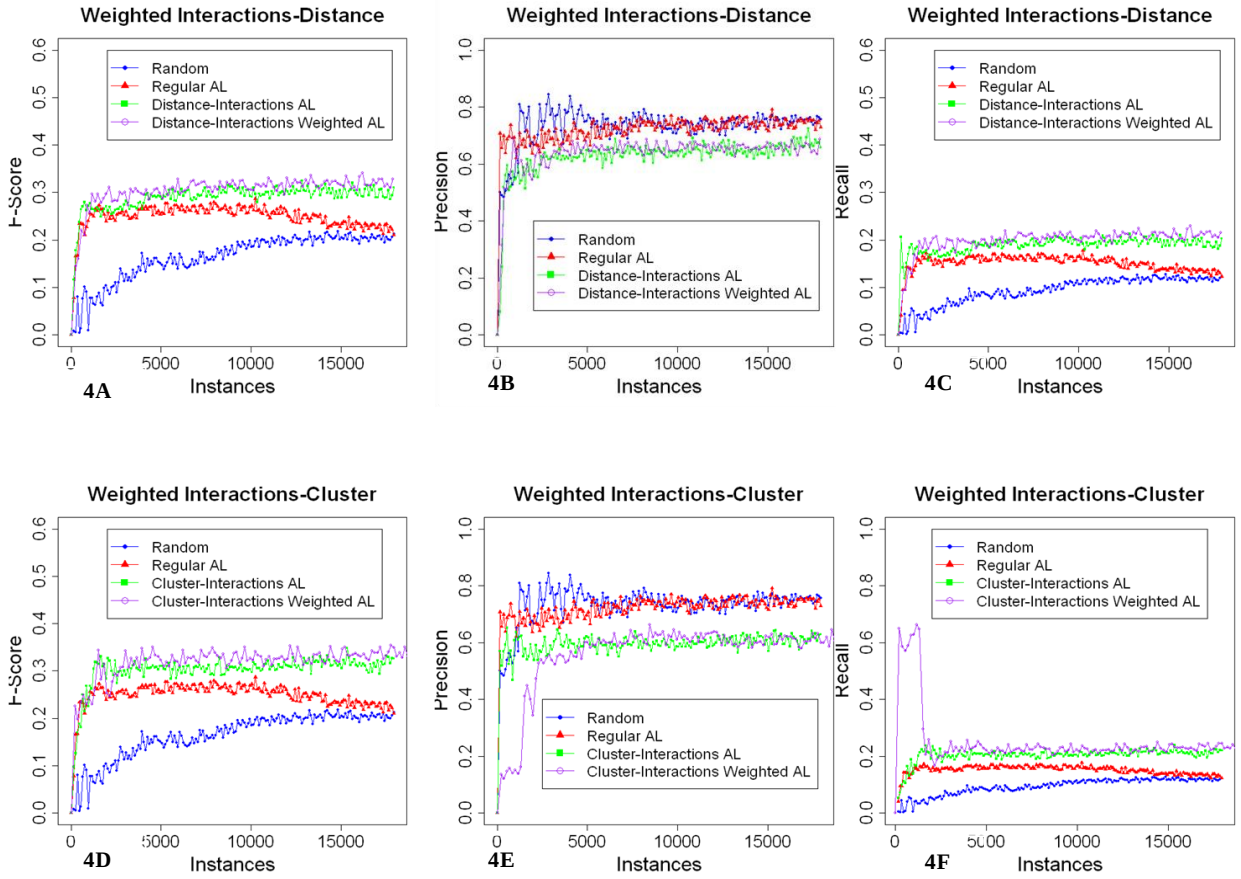


Figure 4 – Results of weighted dataset methods: Performance of weighted-dataset based one-class methods, compared with their counterparts who estimate the labels associated with unlabelled data-points, baseline active learning and random selection of instances. A, B and C show the precision, recall and the F-scores associated with distance-interactions weighted method, compared with distance interactions based methods, regular active learning and random. D, E and F show the precision, recall and the F-scores associated with cluster-interactions weighted method, compared with cluster interactions based methods, regular active learning and the random baseline.

technology for the study of PPIs. Even with HT, it is not feasible to study every molecule or protein-pair. Computational methods are therefore required to predict the annotations. Given the availability of computational methods and HT techniques, it is desirable to have active learning algorithms that guide the selection of some data which when labeled by HT methods improves the accuracy and confidence of labeling the remaining data with computational methods.

We proposed one class active learning heuristics in this paper which deal with a one class reluctant oracle for the domain of protein-protein interaction prediction. They estimate the potential label associated with shortlisted data-points during the active learning process for which the oracle does not give the label. These one class methods have a higher recall and F-score as compared to the traditional active learning counterparts. There is a slight loss in precision, but this value of precision is in itself an underestimate as there is no gold standard dataset for evaluating protein interactions, and false positives in the test set may actually turn out to be true undiscovered interactions.

Although similar to the problem of learning from positive and unlabeled data as in other domains (e.g. document classification [18-20]), PPI discovery is a rare-category problem with large instance space and very small set of highly related features. Traditional approaches to dealing with positive and unlabeled data have already been applied to PPI prediction, and the goal of this work is to discover more interactions beyond those that are already predicted by treating the problem as positive and unlabeled data. Previous methods have

treated learning from positive and unlabeled data as standard. By comparing the methods proposed here with regular active learning and random selection of training data, we show that estimating the label for some of the training data provides superior results.

While proposing these heuristics, we made certain assumptions associated with the underlying domain. We assume that the underlying PPI feature space can characterize oracle behavior; and that the oracle should provide labels that make up a randomly selected subset of the entire interaction space to uncover all potential undiscovered interactions. Although these assumptions may not hold in certain circumstances, given that oracle characterization for PPI prediction is a non-trivial problem, we believe that the one class heuristics proposed in this paper are the stepping stone for solving this problem.

These methods help to incorporate semi-supervised learning approaches to extend active learning to more realistic scenarios by estimating the labels of points for which the oracle does not give the label. These approaches improve the recall of the prediction system.

By characterizing the oracle behavior, we are proposing to use it to exploit the strengths of any experimental method and present an approach for rapid development of the interactome. This would provide a sound basis for rapid discovery of interactomes under budget constraints. Collaborations are underway towards labeling selected instances with yeast 2-hybrid methods.

6. ACKNOWLEDGEMENTS

Funding: MG and SA's work has been funded by the Biobehavioral Research Awards for Innovative New Scientists (BRAINS) grant R01-MH094564 awarded to MG by the National Institute of Mental Health of National Institutes of Health (NIMH/NIH) of USA.

REFERENCES

1. Stumpf, M.P., et al., *Estimating the size of the human interactome*. Proc Natl Acad Sci U S A, 2008. **105**(19): p. 6959-64.
2. Qi, Y., Z. Bar-Joseph, and J. Klein-Seetharaman, *Evaluation of different biological data and computational classification methods for use in protein interaction prediction*. Proteins, 2006. **63**(3): p. 490-500.
3. Hart, G.T., A.K. Ramani, and E.M. Marcotte, *How complete are current yeast and human protein-interaction networks?* Genome Biol, 2006. **7**(11): p. 120.
4. Fields, S. and O. Song, *A novel genetic system to detect protein-protein interactions*. Nature, 1989. **340**(6230): p. 245-6.
5. Ho, Y., et al., *Systematic identification of protein complexes in Saccharomyces cerevisiae by mass spectrometry*. Nature, 2002. **415**(6868): p. 180-3.
6. Settles, B., *Active Learning Literature Survey*. Computer Sciences Technical Report 1648, 2009.
7. Nguyen, H. and A. Smeulders, *Active Learning using Pre-clustering*. International Conference on Machine Learning (ICML): 2004, 2004: p. 623 - 630.
8. Campbell, C., N. Cristianini, and A. Smola, *Query Learning with Large Margin Classifiers*. Proceedings of the Seventeenth International Conference on Machine Learning (ICML-2000), 2000.
9. Pinar Donmez, J. and P. Bennett, *Dual Strategy Active Learning*. Proceedings of the 18th European conference on Machine Learning. Warsaw, Poland, 2007.
10. Mohamed, T., J. Carbonell, and M. Ganapathiraju, *Active learning for human protein-protein interaction prediction*. BMC Bioinformatics, 2010. **11**(Suppl 1): p. S57.
11. Chen, Y. and D. Xu, *Computational analyses of high-throughput protein-protein interaction data*. Curr Protein Pept Sci, 2003. **4**(3): p. 159-81.
12. Donmez, P. and J.G. Carbonell. *Proactive learning: cost-sensitive active learning with multiple imperfect oracles*. in *Proceedings of the 17th ACM conference on Information and knowledge management*. 2008. ACM.
13. Keshava Prasad, T.S., et al., *Human Protein Reference Database--2009 update*. Nucleic Acids Res, 2009. **37**(Database issue): p. D767-72.
14. Venkatesan, K., et al., *An empirical framework for binary interactome mapping*. Nat Methods, 2009. **6**(1): p. 83-90.
15. Elkan, C. and K. Noto. *Learning classifiers from only positive and unlabeled data*. in *Proceeding of the 14th ACM SIGKDD international conference on Knowledge discovery and data mining*. 2008. Las Vegas, Nevada, USA: ACM.
16. Jansen, R., et al., *A Bayesian networks approach for predicting protein-protein interactions from genomic data*. Science, 2003. **302**(5644): p. 449-53.
17. Lin, N., et al., *Information assessment on predicting protein-protein interactions*. BMC Bioinformatics, 2004. **5**: p. 154.
18. Li, X. and B. Liu. *Learning to classify texts using positive and unlabeled data*. in *International joint Conference on Artificial Intelligence*. 2003. LAWRENCE ERLBAUM ASSOCIATES LTD.
19. Liu, B., et al. *Partially supervised classification of text documents*. in *MACHINE LEARNING-INTERNATIONAL WORKSHOP THEN CONFERENCE-*. 2002. Citeseer.
20. Yu, H., J. Han, and K.C.C. Chang, *PEBL: Web page classification without negative examples*. Knowledge and Data Engineering, IEEE Transactions on, 2004. **16**(1): p. 70-81.

LOOKING FOR A MISSING LINK IN THE NETWORK

Laura Astola[1,2]*, Simon van Mourik*[1,2] *and Jaap Molenaar*[1,2]

[1]Biometris, Wageningen University and Research Centre,
P.O. Box 100, 6700 AC Wageningen, The Netherlands
[2]Netherlands Consortium for Systems Biology, Amsterdam, The Netherlands,
laura.astola@wur.nl, simon.vanmourik@wur.nl, jaap.molenaar@wur.nl

ABSTRACT

We consider the problem of inferring a biological network from given experimental data. In earlier work on metabolic pathway inference, we studied a situation where the molecular interactions were fairly well known, the kinetic parameters were completely unknown and the network topology was almost known. Starting from a similar setting, using simulations, we have investigated the inference of missing links in a network. We use two approaches, a deterministic one using ordinary differential equations and one employing a statistical approach. We find that the deterministic fit-based approach yields quite positive results in linear as well as in a non-linear context.

INTRODUCTION

The motivation for this study rose from our previous work [1–3]. The objective was to find those enzymes that were responsible for the glycosylation of flavonoids in tomatoes. The topology of the metabolic network was known up to one edge, since it was not known whether an (unknown) enzyme can attach two or more molecules at once to a precursor of a flavonoid. This lead us to investigate methods for finding a missing edge in biological networks. A network can be represented as a graph, where the nodes correspond to the measured components, such as metabolite concentrations, and the edges imply interactions between the components. If the topology, i.e., the network structure, is fixed, one may estimate the unknown parameters to fit the data. A common approach is to choose a combination of topology and parameters that best explains the data in the sense of maximum likelihood. These network topologies need not always be inferred from scratch, since there is often preliminary knowledge available. However, when several putative network scenarios are present, we are faced with a model discrimination challenge: how to exploit the measurements to infer the correct model? It is well known that a network inference problem is identifiable (tractable) when the network has a tree-like structure [1, 4]. To expand such a tree structure to an optimal network, adding edges one by one may well be a successful methodology [5]. In this procedure it is highly relevant to detect a missing edge at each stage of inference. Although, in general, there is a wide variety of methodologies for biological network reconstruction described in the literature, including Boolean networks, association-based networks such as co-expression analysis, deterministic or stochastic ordinary differential equations (ODEs) and graphical models, reliable network inference remains elusive [6–9]. In our present study, we investigate whether we can detect a missing edge in both a deterministic and a statistical context. We used two types of simulated data: linear ODEs under mass-balance constraint for kinetic networks and non-linear ODEs based on the model for regulatory networks in the DREAM challenge [10]. In the following section we explain our fit-based heuristic approach. Then we discuss the alternative statistical model approach that we employed in parallel. In subsequent sections we specify how we generated the simulation data, show the inference results with the two methodologies and finish with some conclusions.

DETERMINISTIC APPROACH

In an deterministic approach we have at our disposal a certain (linear, or non-linear) model of the network, that can be represented by a system of ODEs. If the network topology would be fully known, we could use this model and optimize the parameters to fit the data. However, we consider a restricted problem, where the network topology is almost known, except for one missing edge. A heuristic approach is to take the known part (incomplete) network, estimate the parameters as if the topology was correct and then inspect the obtained fit against measurement data. If some part of the data is relatively well reconstructed while another part is significantly less well fit, we may suspect that the missing edge connects variables within the latter set. Based on this straightforward line of reasoning we test the following "fit-based"-scheme:

1. Fit the incomplete network to data and estimate the parameters, using, e.g., NMinimize (Mathematica) or fmincon(matlab) and record the residuals for each variable;

2. Select the two variables with largest residuals as candidate nodes that share the missing edge.

Such an approach is not likely to work when many edges are missing or dislocated, but here we test how well it can detect a single missing link. We applied this approach to

linear as well as non-linear models and discuss our findings in the RESULTS section.

STATISTICAL APPROACH

Among the statistical network inference methods, graphical models are currently quite popular. Graphical models usually refer to a combination of graph theory and probabilistic reasoning [4]. In terms of a network graph, the nodes represent the same variables, (metabolite concentrations etc.), as in the deterministic case. However, the edges, instead of bio-mechanical interactions, now code for causal dependencies between the variables. Just as in the deterministic case, one wants to find a combination of network topology and parameters that corresponds to maximum likelihood of observing the data. In the probabilistic context this is achieved by maximizing the probability of the underlying network using Bayes' rule:

$$P(G|D,\theta) \propto P(D|\theta,G)P(G)\ , \tag{1}$$

where G is the network topology, D are the data, and θ the parameters of the probability distributions. To compute the probability over all networks usually implies that one has to resort to Markov Chain Monte Carlo sampling of networks and parameters. When the network is small, one may apply Bayes' rule and exploit the conditional independencies to make the inference tractable. On the other hand, when the conditional independencies between the variables are known, we also know the graph of G and can further estimate the variables θ in $P(D|\theta,G)$. In the literature numerous approaches have been proposed to compute these conditional independencies empirically from data [11]. However, since we here consider only the specific case of one missing link, we do not need this whole machinery for the inference. For each variable X_i, we want to find a candidate variable X_j, that is most likely to improve the fit of X_i when an edge between X_i and X_j is added to the network. To compute such causal dependencies of variables with time series data, we adopt the rather simple and intuitive Granger-causality [12]. According to this, if a signal $X_j(t)$ "Granger-causes" a signal $X_i(t)$, then past values of X_j should contain information that helps predict X_i better than the information contained in past values of X_i alone. In practice such a pair-wise causality can be computed by comparing the residuals R_i and R_i^* in the following two expressions, where the parameters A_{ijk} are optimized to fit the data for X_i, where i runs over the dependent variables (regressand), j is some independent variable (regressor) and k indicates successive time lags. The maximum time lag we take into consideration is l.

$$\begin{aligned} X_i(t) &= \sum_{k=1}^{l} A_{iik}X_i(t-k) + R_i(t) \\ X_i(t) &= \sum_{k=1}^{l} A_{iik}X_i(t-k) + \sum_{k=1}^{l} A_{ijk}X_j(t-k) + R_i^*(t) \end{aligned} \tag{2}$$

If the variance of the residuals $R_i^*(t)$ is smaller than that of $R_i(t)$, X_j is said to Granger-cause X_i. We modify this philosophy to serve our particular problem of finding a missing edge and set up the following Algorithm 1 to discriminate between edge candidates. For compact presentation, we denote by $\mathcal{L}_i(X_j,t)$ the inclusion of variable X_j to predict variable X_i.

$$\mathcal{L}_i(X_j,t) = \sum_{k=1}^{l} A_{iik}X_i(t-k) + \sum_{k=1}^{l} A_{ijk}X_j(t-k) + R_i^*(t) \tag{3}$$

$\mathcal{L}_i(X_j,t)$ may take varying number of arguments, for example when expressing variable X_2 as a time-lagged linear combination of itself and variables X_1 and X_5:

$$\begin{aligned} \mathcal{L}_2(X_1,X_5,t) = &\sum_{k=1}^{l} A_{22k}X_2(t-k) + \sum_{k=1}^{l} A_{21k}X_1(t-k) \\ &+ \sum_{k=1}^{l} A_{25k}X_5(t-k) + R_2^*(t) \end{aligned} \tag{4}$$

Algorithm 1 Select the endpoints (m, s_m) of the missing edge

for $i = 1$ **to** #(variables) **do**
  neighbours$_i$ = variables connected to X_i
  foreigners$_i$ = variables not connected to X_i
  for $k = 1$ **to** #(foreigners$_i$) **do**
    $R_{ik}^* =$
    $\sum_{t=0}^{T}(X_i(t) - \mathcal{L}_i(\text{neighbours}_i, \text{foreigners}_i(k), t))^2$
  end for
  $s_i = \arg\min_k R_{ik}^*$
end for
$m = \arg\min_i R_{i,s_i}^*$

In the inference we used a time lag of 3, but in our simulations varying the lag did not change the results significantly. We show the outcomes of applying this algorithm to simulated data in the RESULTS section.

SIMULATIONS

For the linear networks we simulated connected graphs, that have $5,\ldots,12$ nodes and $5,\ldots,15$ edges. For each such node-edge pair we generated 100 random linear ODE-systems. We took 11 samples at equally spaced time-intervals as noiseless data and added 10% Gaussian noise for the noisy data. All parameters in the simulation were chosen to have the same order of magnitude. As an example a simulated random network with 5 nodes, 5 edges, and the corresponding concentration data are shown in Figure. 1. For the non-linear simulations we used the 11 node regulatory network model from the "network topology and parameter estimation challenge" of the DREAM project [10], where repression and activation of gene expression were modeled using Hill-type functions. For example, the rate of change of protein $p1$ concentration that

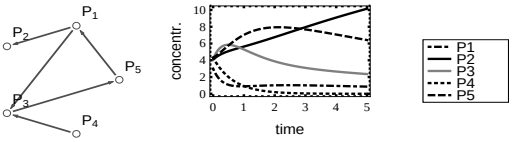


Figure 1. An example of a simulated graph with 5 nodes and 5 edges and the corresponding dynamics in the linear case.

depends on the concentrations of proteins $p2$ (activator) and $p3$ (repressor) can be modeled as follows:

$$\frac{dp1(t)}{dt} = \frac{sp1 \cdot (\frac{p2(t)}{r1_{Kd}})^{r1_h}}{(1+(\frac{p2(t)}{r1_{Kd}})^{r1_h})(1+(\frac{p3(t)}{r2_{Kd}})^{r2_h})} - dp1 \cdot p1(t)\ , \quad (5)$$

where sp_i denotes the synthesis rate, dp_i the degradation rate of protein p_i and ri_{Kd} the dissociation constant of the reaction and ri_h the Hill-coefficient (of interaction). The original model of DREAM project contained 55 unknown parameters, but to do inference in reasonable time scale, we fixed most of the parameter values and inferred only 11 variables, for example the protein synthesis rates, under different experimental conditions, where one of the genes is overexpressed. We collected data from 7 experiments, one as reference wild type-data and the rest where genes $1, 2, 3, 5, 7$, and 11 were overexpressed (cf. Figure 3).

RESULTS

In case of the linear model, we did 3100 reconstructions using the fit-based heuristics and Algorithm 1, which we compared to results from random guessing. In case of random guessing, the existing edges (that cannot correspond to a missing edge) were removed from the candidate pool, whereas in the fit-based and Granger-based approaches we did not make use of this knowledge and focused only on the residuals. The results show that a simple fit-based approach is most successful to detect a single missing link (cf. Figure. 2). For the non-linear case we experimented only with the fit-based heuristic approach, by removing an edge that is forming a repressing feedback loop from node $p7$ to node $p1$. We did not modify the original parameters too much, so as not to lose the delicate oscillatory behavior. An illustration of the missing edge as well as the results are shown in Figure. 3. In the graphics at the bottom right, we see peaks at nodes $p1$ and $p7$ as expected. However, if the variables $p2$ or $p3$ are, say, 5 times higher overexpressed, this results in higher residuals for variables $p2$ and $p3$ in experiments 2 and 3, although they are not part of the missing edge. This is due to the fact that concentration changes in $p1$ have immediate effect on the concentrations of $p2$ and further of $p3$ via activation.

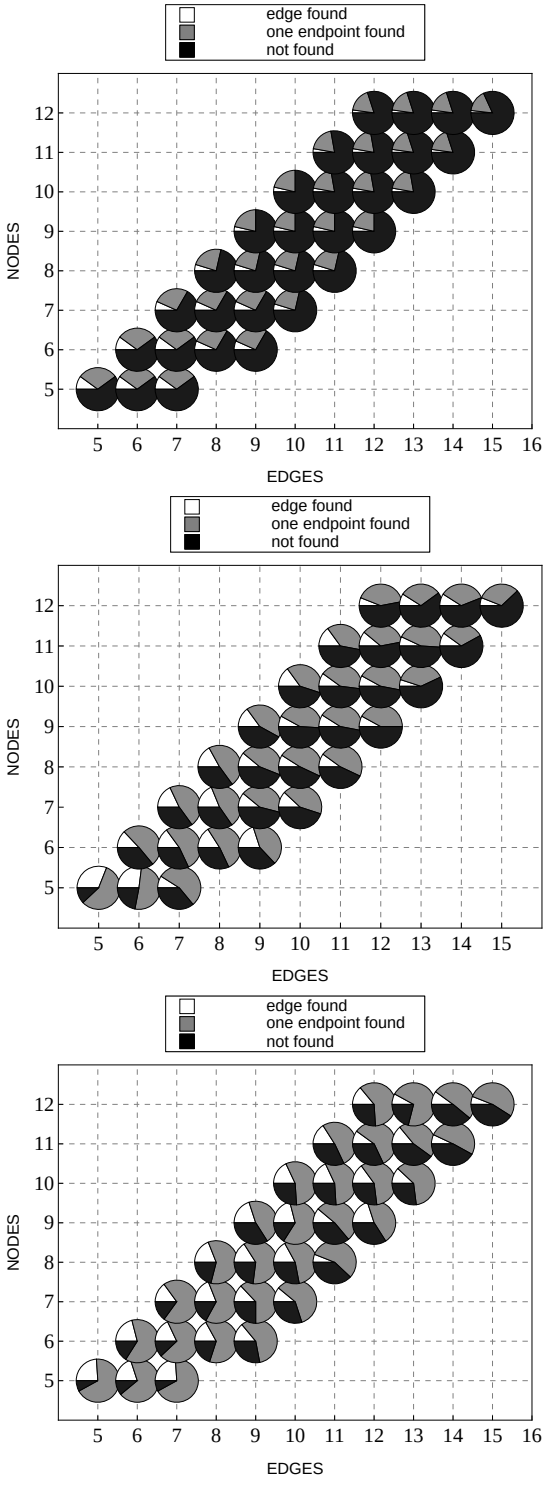


Figure 2. Results for the linear simulations. We have plotted the proportions of three complementary cases: a) the correct missing edge was found, b) at least one endpoint of the missing edge was found, c) none of these were found. Top: By pure guessing. Middle: Using Granger-causality. Bottom: Using fit-based heuristics. The fit-based inference of a missing link is here the most successful approach.

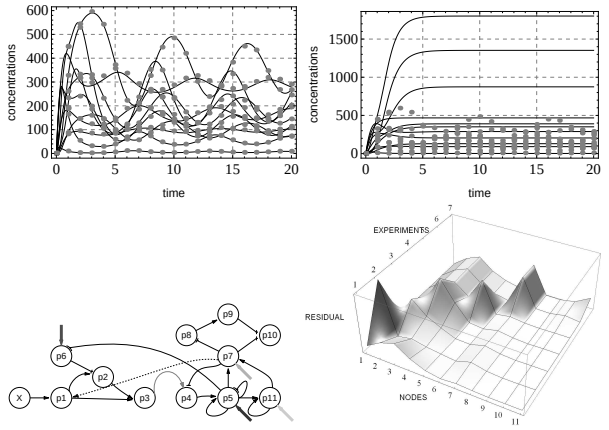


Figure 3. Results for the non-linear simulations. Top left: Dots are data, lines give the dynamics of the system if the parameters are fitted using correct topology. Top right: As on left hand side, but in case parameters were fitted with a network topology missing one edge. Bottom left: The structure of the model network, the missing edge is indicated with a dotted line. Bottom right: The residuals of fitting the variables to data per experiment. Although the results are discrete, we used interpolation for visual continuity.

CONCLUSIONS

Despite the Granger-causality being more rigorous and versatile than the simple fit-based inference method, the latter still seems to better capture the missing edges (cf. Figure 2). Since the fit-based heuristics is also based on a linear model, this can partly explain the better performance. To reduce such potential effect, we compared also the results in case the data are noisy. By using both, noiseless and noisy data and counting the cases, where the fit-based method performed better, we obtained:

- (noiseless) fit-based, (noiseless) Granger: 100% ;
- (noisy) fit-based, (noisy) Granger: 100% ;
- (noisy) fit-based, (noiseless) Granger: $61,3\%$.

In the non-linear case, the residuals were also significantly larger for variables that constitute the missing link. In our experiment, a repressor of variable $p1$ was removed, resulting in higher expression of $p1$. Variables $p2$ and $p3$ being connected to $p1$ in cascade were then also indirectly upregulated. Therefore, overexpression of $p2$ and $p3$ can result in higher residuals at nodes $p2, p3$ than at the nodes $p1$ and $p7$ that correspond to the actual missing link. However, in general, based on our simulation experiments we found that the intuitive fit-based heuristics quite often correctly points towards the missing link. The computation time of the fit-based method grows almost linearly w.r.t. increasing number of nodes, whereas the Granger-based method shows exponential growth making the extension to large scale networks impractical.

1. REFERENCES

[1] L. Astola, M. Groenenboom, V. Gomes Roldan, F. Eeuwijk, R. Hall, A. Bovy, and J. Molenaar, "Metabolic pathway inference from time series data: a non iterative approach," in *Lecture Notes in Bioinformatics*. 2011, vol. 7036, pp. 97–108, Springer.

[2] L. Astola, V. Gomes Roldan, and J. Molenaar, "Inferring the genes underlying flavonoid production in tomato," in *9th International Workshop on Computational Systems Biology*, Ulm, Germany, June 2012.

[3] L. Astola, V. Gomes Roldan, F. van Eeuwijk, R. D. Hall, M. Groenenboom, and J. Molenaar, "Tree graphs and identifiable parameter estimation in metabolic networks," *Submitted to BMC Systems Biology*, 2012.

[4] D. Barber, *Bayesian Reasoning and Machine Learning*, Cambridge University Press, 2012.

[5] D. Heckerman, D. Geiger, and D. Chickering, "Learning Bayesian networks: the combination of knowledge and statistical data," *Machine Learning*, vol. 20, no. 3, pp. 197–243, 1995.

[6] P. D'haeseleer, S. Liang, and R. Somogyi, "Genetic network inference: from co-expression clustering to reverse engineering," *Bioinformatics*, vol. 16, no. 8, pp. 707–726, 2000.

[7] H. Lähdesmäki and I. Shmulevich, "Learning the structure of dynamic Bayesian networks from time series and steady state measurements," *Machine Learning*, vol. 71, pp. 185–217, 2008.

[8] C. A. Penfold and D. L. Wild, "How to infer gene networks from expression profiles, revisited," *Interface Focus*, vol. 1, no. 6, pp. 857–870, December 2011.

[9] D. Marbach, R. Prill, T. Schaffter, C. Mattiussi, D. Floreano, and G. Stolovitzky, "Revealing strengths and weaknesses of methods for gene network inference," *PNAS*, vol. 107, no. 14, pp. 6286–6291, 2010.

[10] Sauro, H. and Meyer, P. and Saez Rodriguez, J. and Stolovitzky, G. and Chandran, D. and Kim, K. H. and Cokelaer, T., "DREAM7, Network topology and parameter inference challenge," `http://www.the-dream-project.org/`, 2012.

[11] Zhang, K. and peters, J. and Janzing, D. and Schoelkopf, B., "Preprint: Kernel-based Conditional Independence Test and Application in Causal Discovery," `http://arxiv.org/abs/1202.3775`, 2012.

[12] C. Granger, "Investigating causal relations by econometric models and cross-spectral methods.," *Econometrica*, vol. 37, no. 3, pp. 424–38, 1969.

CHARACTERIZATION AND IDENTIFICATION OF TISSUE-SPECIFIC PROMOTERS IN RICE

Wen-Chi Chang[1], Mou-Yuan Yeh[2], Tzong-Yi Lee[3], and Ying-Chi Wen [1]

[1] Institute of Tropical Plant Sciences, National Cheng Kung University,
[2] Department of Computer Science and Information Engineering, National Cheng Kung University,
[3] Department of Computer Science and Engineering, Yuan Ze University,
No.1, University Road, Tainan City 70101, Institute of Tropical Plant Sciences, Taiwan,
sarah321@mail.ncku.edu.tw, ismike2000@hotmail.com, francis@saturn.yzu.edu.tw,
bonble@gmail.com

ABSTRACT

Background: A set of genes in a particular tissue might have similar expression profiles, as well as involved in similar functions during different plant developmental stages. As is widely assumed, such tissue-specific genes are regulated by a set of similar TFs, and their promoters contain similar regulatory patterns. While receiving considerable attention in animals, this topic has seldom been examined in plants, especially rice. Therefore, this study thoroughly elucidates promoter features of tissue-specific genes in rice by combining computational methods with experimental data.

Results: Tissue-specific genes are identified using microarray data involving different stages of the development of rice. Genes with Z-scores that exceed a threshold value (Z score> 2.5) are defined as tissue (organ)-specific genes. The total number of genes thus identified is 1744. In Gene Ontology (GO) enrichment analysis, most tissue (organ)-specific genes respond to stress and stimulus. Furthermore, more CpG/CpNpG islands were detected in non-tissue (organ)-specific promoters. The base composition profiles of tissue (organ)-specific promoters prefer C over G in the downstream region near transcription start sites (TSSs), yet have no preference in non-tissue-specific promoters. Finally, tissue-specific structures in the gene promoters are determined using motif search methods. Several novel tissue (organ)-specific motifs and known TFBSs are identified, including CDC5 and BZR1.These tissue-specific motifs may play a prominent role in the development of rice.

Conclusions: Based the results of this study, several important tissue (organ)-specific features are recognized in each tissue. This study first elucidates the components of promoters that distinguish nonspecific from tissue (organ)-specific genes in rice. These analysis methods can help us to understand the mechanism of transcription regulation under various developmental stages.

1. INTRODUCTION

The regulation of gene expression is dynamic across various developmental stages in plants. Some genes are expressed at a special time, in a specific tissue, or under particular conditions. Hence, identifying a set of genes that are expressed under a specific time point or tissue (organ) is important for understanding the morphology and physiology of a tissue or organ systems. DNA microarray high-throughput technique has been widely applied in the recent decade to examine transcriptional expression patterns on the whole-genome scale. By using this approach, numerous studies have investigated dynamic gene expression profiles in various developmental processes in plants [1-3]. For instance, a previous study established a gene expression map of *Arabidopsis thaliana* development by using various developmental samples via the microarray high-throughput method [4], subsequently providing valuable information about which gene group is critical in which developmental stages or tissues. Wang *et al*. identified 2667 microarray probe sets that were expressed differentially in four stages of panicle development, indicating that RFL and LAX have an essential role for determining the inflorescence architecture in rice [3]. Additionally, transcription factors (TFs) and their corresponding *cis*-acting elements in promoters have received considerable interest in gene regulation research [5]. Therefore, thoroughly elucidating TFs and their binding sites in promoters is essential to studying the regulation of transcription.

To advance knowledge of co-expressed gene regulation in plant sciences, related studies have developed for investigating co-occurrence transcription factor binding sites (TFBSs) in a group of gene promoters. PlantPAN [6], a database- assisted system, analyzes the co-occurrence of combinatorial TFBSs with a distance constraint in plant co-expressed genes. ATTED-II [7] provides co-regulated genes based on the co-expressions of genes that are deduced from microarray data and predicted *cis*-regulatory elements in their upstream sequence. AtPAN [8] provides an integrated system for

reconstructing transcriptional regulatory networks based on microarray co-expression data in *Arabidopsis*. Haberer*et al.* utilized a comparative genomics approach to determine the conservation motifs in a group of co-expressed gene promoters that belong to a specific biochemical pathway [9].Walther *et al.* analyzed large-scale properties of promoters and found highly significant positive correlations between the density of *cis*-elements in the promoters and the number of conditions under which a gene is regulated differentially [10]. Despite the numerous studies to identify the co-occurrence regulatory motifs in co-expressed gene promoters, the tissue-specific structure patterns in plant promoters have seldom been investigated. Moreover, related research tends to focus mainly on *Arabidopsis*, rarely examining other plants. As an important global food crop, rice is a model for genomic research on cereals. Surprisingly, the issue of interest herein has never been large-scale examined in relation to rice. Therefore, this study elucidates the tissue-specific structure patterns in rice promoters.

By combining computational methods with experimental data, this study thoroughly analyzes promoter features that are related to tissue (organ)-specific genes in rice. Tissue (organ)-specific expression genes are first identified using microarray data from different stages of development of rice. Tissue-specific structures in those gene promoters are then identified as well by using many motif search methods. Based on those results, numerous unknown motifs can be found in the promoters of tissue (organ)-specific genes. These tissue (organ)-specific motifs may be novel TFBSs that profoundly impact rice development.

2. GENERAL INSTRUCTIONS

Figure 1 schematically depicts the flow chart of this study. The microarray expression data of rice was first downloaded from a public database and preprocessed to identify tissue-specific genes. The promoter sequences were then extracted from the RGAP database (Rice Genome Annotation Project, http://rice.plantbiology.msu.edu/). Following determination of the promoter region, the known TFBSs and *de novo* conserved motifs among the same group of promoters were annotated. The redundant motifs were subsequently removed and, in doing so, tissue-specific structural patterns in the rice promoter could be defined. Details of the above methods are described as follows.

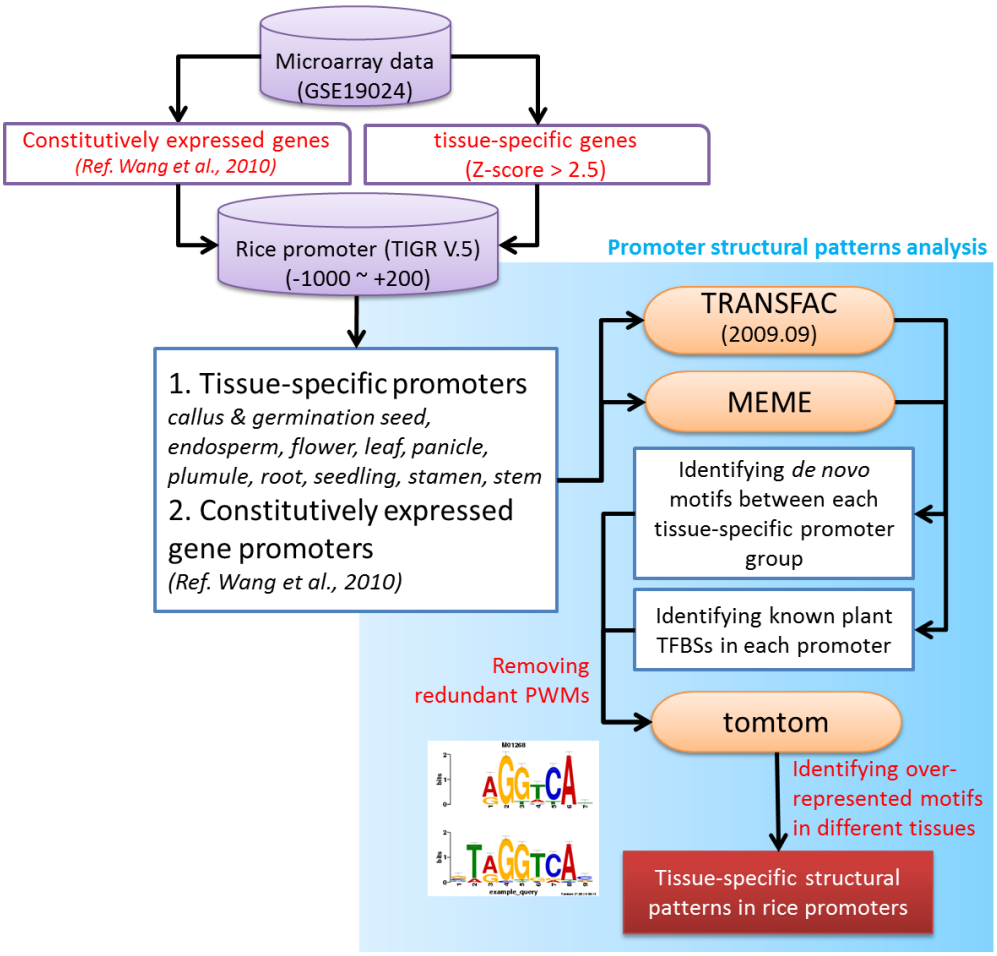


Figure1 System flow of this study

2.1. Tissue-specific datasets and GO analysis

Raw microarray expression data were obtained from the NCBI Gene Expression Omnibus (GEO) database. The dataset of Wang *et al.* (GSE19024) [3], which includes transcriptional expression data for 39 tissues/organs from rice, was used. The expression values of all samples were preprocessed and normalized using the Affy-package [11], as made available in the Bioconductor software suite in the R statistical analysis program. The normalization method was RMA. Based on the transcriptome analysis of Wang *et al.*[3], tissues (organs) were grouped into ten clusters. The ten clusters of tissues (organs) were germination seeds, endosperm, plumule, seedling, root, stem, leaf, flower, stamen, and panicle. For each gene, normalized expression data were converted into Z-scores by using the mean and the standard deviation of gene's expression values. Refer to previous studied [12, 13], when the Z-score exceeded the threshold value of 2.5, the corresponding genes were regarded as "tissue (organ)-specific", the rest were nonspecific. To determine whether genes differentially expressed in specific tissues (organs) belonged to particular Gene Ontology (GO) categories, the GOSlim assignment of rice loci was downloaded from the RGAP database [14]. The GOSlim terms were obtained from the GO Database [15]. Finally, GO term enrichment analysis of each gene was performed using a homemade PHP program.

2.2. CpG islands and analysis of base composition

Although CpG islands with tissue-specific genes in mammals have received considerable attention recently [16], plants have received lesser attention in this respect. CpG island searcher is a program for identifying CpG/CpNpG islands [5].The CpG/CpNpG islands are defined as those DNA regions that are longer than "window size" (200 bp or 500 bp), with a moving average C+C frequency of over "GC percentage (GC%)" (i.e. the value of GC% used in this study includes 50%, 55%, 60%, 65%, and 70%) and a moving average CpG/CpNpG observed/expected (o/e) ratio over "OE CpG" (the value of OE CpG used in this study includes 0.6, 0.65, 0.7, 0.75, 0.8, 0.85, 0.9, and 1.0). CpG/CpNpG islands in rice gene promoters were identified using every combination of parameters. Additionally, the base composition profiles with 10 bp window size were calculated for ubiquitous and tissue-specific genes with and without CpG/CpNpG islands.

2.3. Identifying tissue (organ)-specific motif

The tissue (organ)-specific motifs in tissue (organ)-specific gene promoters were identified in a two-phase study. Multiple EM for Motif Elicitation (MEME) is an extensively used approach for finding novel 'signals' in sets of biological sequences [17]. Firstly, *de novo* common sequences were identified separately in ten clusters of tissue (organ)-specific promoters by using the MEME program. The minimum and maximum motif widths were set to 4 and 15, respectively, in the MEME program. As an effective means of evaluating the similarity between two DNA motifs, the TOMTOM software program [18] compares a query DNA motif with the motifs in a database of known motifs and then ranks the matching motifs by p-value and q-value. The p-value is related to the minimum number of overlapping positions in a given offset. The q-value is the minimum false discovery rate at which

the observed similarity is deemed significant. Whenever p-value $\leqq$ 0.001 and q-value $\leqq$ 0.05 for a particular pair of DNA motifs (PWMs) in the TOMTOM software program, both motifs were regarded as similar (redundant) ones. The redundant PWMs were removed from each cluster before determining whether the motif is associated with novel or known TFBSs. All plant position weight matrices (PWMs) from TRANSFAC database version 2009.9 [4] were used to populate a database of known TFBSs for comparison by TOMTOM. Consequently, the novel and known TFBSs in tissue (organ)-specific gene promoters can be identified. Finally, the motif logos were generated using the WebLogo program [19].

3. RESULTS AND DISCUSSION

3.1. Identification of tissue (organ)-specific genes and functional analysis

A total number of 1744 genes were identified as tissue (organ)-specific genes after analysis (Table 1). A GO enrichment analysis was performed of GO terms in 1744 tissue (organ)-specific genes. In the biological process, most tissue (organ)-specific genes respond to stress, especially in a germination seed, leaf, panicle, plumule, root, seedling, and stamen (Fig. 2A and Table2). Moreover, over 20% of tissue (organ)-specific genes are descendants of the GO term "response to stimulus". This term refers to "response to endogenous stimulus", "response to abiotic stimulus", or "response to biotic stimulus". Above findings are similar to those of previous studies, which have demonstrated that genes according

Table 1 The statistics of tissue-specific genes in rice. (Z-score>2.5)

Organs(Tissues)	No. of genes
germination seeds	97
endosperm	232
plumule	17
seedling	39
root	122
stem	25
leaf	123
flower	36
stamen	925
panicle	128
Total	1744

to tissue-specific genes are responsive to stimuli [16]. In the cellular component, 22.14% of tissue (organ)-specific genes are located in the plasma membrane (Fig. 2B). These results are similar to biological process analysis, implying that those genes are essential to the response of stress and stimulation in the

Table 2 GO analysis of tissue-specific genes

Gene Ontology \ Tissue	germination seeds	endosperm	plumule	seedling	root	stem	leaf	flower	stamen	panicle	Total
Biological Process											
signal transduction	5.56%	8.40%	0.00%	6.90%	4.76%	0.00%	15.28%	10.20%	8.39%	3.61%	8.40%
cell death	0.93%	0.00%	0.00%	1.72%	0.68%	0.00%	0.87%	0.00%	0.51%	0.00%	0.56%
cell differentiation	0.93%	0.00%	0.00%	0.00%	1.36%	4.17%	3.93%	2.04%	0.64%	1.20%	1.24%
cell cycle	0.93%	0.84%	0.00%	1.72%	0.00%	0.00%	0.00%	0.00%	0.25%	0.00%	0.31%
cell growth	0.93%	0.00%	0.00%	0.00%	2.72%	8.33%	0.87%	0.00%	1.14%	0.00%	1.11%
cell organization and biogenesis	4.63%	4.20%	0.00%	0.00%	4.08%	0.00%	0.44%	2.04%	8.26%	6.02%	5.44%
cell homeostasis	0.00%	0.00%	0.00%	0.00%	0.68%	0.00%	0.00%	0.00%	1.40%	0.00%	0.74%
transport	3.70%	5.88%	6.67%	3.45%	6.80%	**12.50%**	1.31%	4.08%	6.86%	3.61%	5.50%
embryonic development	0.00%	0.84%	0.00%	0.00%	0.00%	0.00%	0.44%	0.00%	0.25%	0.00%	0.25%
flower development	1.85%	0.84%	0.00%	1.72%	0.00%	0.00%	1.31%	0.00%	1.14%	2.41%	1.11%
morphogenesis	0.00%	0.00%	0.00%	0.00%	0.00%	0.00%	0.87%	0.00%	0.25%	1.20%	0.31%
regulation of gene expression, epigenetic	0.00%	0.84%	0.00%	0.00%	0.00%	0.00%	0.00%	0.00%	0.00%	2.41%	0.19%
growth	0.00%	0.84%	0.00%	1.72%	0.68%	0.00%	1.31%	0.00%	0.38%	0.00%	0.56%
reproduction	0.00%	0.84%	0.00%	0.00%	0.68%	4.17%	0.44%	0.00%	0.76%	2.41%	0.74%
post-pollination	0.00%	0.00%	0.00%	0.00%	0.00%	4.17%	0.87%	0.00%	0.25%	0.00%	0.31%
photosynthesis	0.00%	0.00%	0.00%	1.72%	0.00%	0.00%	0.00%	0.00%	0.00%	0.00%	0.06%
response to stress	**16.67%**	11.76%	**26.67%**	**15.52%**	**20.41%**	8.33%	**17.47%**	10.20%	**12.96%**	**21.69%**	**14.95%**
response to endogenous stimulus	12.04%	9.24%	6.67%	10.34%	16.33%	4.17%	14.85%	8.16%	7.37%	2.41%	9.51%
response to abiotic stimulus	4.63%	4.20%	6.67%	5.17%	4.76%	8.33%	3.93%	2.04%	6.35%	12.05%	5.74%
response to biotic stimulus	8.33%	4.20%	13.33%	1.72%	14.29%	4.17%	6.11%	8.16%	5.34%	4.82%	6.36%
amino acid and derivative metabolism	4.63%	4.20%	6.67%	5.17%	1.36%	8.33%	2.18%	8.16%	3.68%	0.00%	3.46%
biosynthesis	7.41%	6.72%	6.67%	13.79%	2.72%	**12.50%**	3.93%	**16.33%**	3.56%	2.41%	4.88%
carbohydrate metabolism	0.00%	3.36%	0.00%	0.00%	0.68%	0.00%	0.44%	4.08%	2.67%	1.20%	1.85%
catabolism	0.93%	1.68%	6.67%	3.45%	1.36%	0.00%	0.00%	0.00%	1.91%	2.41%	1.54%
energy pathways	0.00%	1.68%	0.00%	0.00%	0.00%	0.00%	0.00%	0.00%	0.38%	0.00%	0.31%
lipid metabolism	5.56%	2.52%	0.00%	6.90%	2.04%	8.33%	2.62%	4.08%	1.40%	0.00%	2.29%
DNA metabolism	0.00%	2.52%	0.00%	0.00%	0.00%	0.00%	0.00%	0.00%	0.64%	2.41%	0.62%
transcription	5.56%	**14.29%**	0.00%	3.45%	4.08%	0.00%	1.75%	4.08%	2.67%	20.48%	4.63%
protein biosynthesis	0.93%	2.52%	6.67%	0.00%	1.36%	0.00%	0.00%	0.00%	12.07%	4.82%	6.55%
protein modification	9.26%	4.20%	6.67%	3.45%	6.80%	0.00%	15.72%	6.12%	7.24%	2.41%	7.78%
secondary metabolism	4.63%	3.36%	6.67%	12.07%	1.36%	12.50%	3.06%	10.20%	1.27%	0.00%	2.72%
Cellular Component											
cell wall	17.86%	14.29%	**40.00%**	9.09%	**29.17%**	0.00%	16.52%	13.33%	11.68%	3.33%	13.87%
cytoskeleton	0.00%	0.00%	0.00%	0.00%	0.00%	0.00%	0.00%	0.00%	2.38%	3.33%	1.52%
cytosol	0.00%	6.12%	0.00%	0.00%	0.00%	16.67%	0.00%	6.67%	13.47%	10.00%	8.86%
endoplasmic reticulum	17.86%	12.24%	20.00%	15.15%	4.17%	16.67%	4.35%	13.33%	5.35%	13.33%	6.88%
endosome	0.00%	0.00%	0.00%	3.03%	0.00%	0.00%	0.00%	0.00%	0.00%	0.00%	0.12%
Golgi apparatus	0.00%	10.20%	0.00%	3.03%	5.56%	0.00%	0.87%	6.67%	4.16%	0.00%	3.85%
mitochondrion	7.14%	12.24%	20.00%	21.21%	19.44%	0.00%	24.35%	26.67%	14.06%	**20.00%**	16.20%
peroxisome	3.57%	0.00%	0.00%	0.00%	0.00%	0.00%	0.00%	0.00%	0.20%	0.00%	0.23%
plastid	7.14%	8.16%	0.00%	**24.24%**	4.17%	16.67%	5.22%	0.00%	5.35%	10.00%	6.29%
ribosome	0.00%	2.04%	20.00%	0.00%	0.00%	0.00%	0.00%	0.00%	**17.03%**	10.00%	10.61%
vacuole	0.00%	6.12%	0.00%	0.00%	6.94%	16.67%	0.00%	0.00%	2.97%	3.33%	2.91%
nuclear membrane	0.00%	2.04%	0.00%	0.00%	0.00%	0.00%	0.87%	6.67%	0.20%	3.33%	0.58%
nucleolus	3.57%	6.12%	0.00%	0.00%	0.00%	0.00%	0.87%	0.00%	3.37%	3.33%	2.68%
nucleoplasm	0.00%	0.00%	0.00%	0.00%	0.00%	0.00%	0.00%	0.00%	0.20%	0.00%	0.12%
thylakoid	3.57%	2.04%	0.00%	15.15%	1.39%	0.00%	1.74%	0.00%	2.97%	6.67%	3.15%
plasma membrane	**39.29%**	**18.37%**	0.00%	9.09%	**29.17%**	**33.33%**	**45.22%**	**26.67%**	16.63%	13.33%	**22.14%**
Molecular Function											
hydrolase	**28.77%**	17.44%	18.18%	4.00%	**19.09%**	11.76%	5.39%	**21.88%**	**19.94%**	12.63%	**17.03%**
kinase	12.33%	4.65%	0.00%	8.00%	10.00%	0.00%	**22.55%**	9.38%	11.33%	5.26%	11.79%
transferase	9.59%	10.47%	**36.36%**	**32.00%**	17.27%	**17.65%**	5.88%	18.75%	8.01%	6.32%	9.66%
enzyme regulator activity	0.00%	2.33%	0.00%	0.00%	0.00%	0.00%	0.00%	0.00%	1.51%	0.00%	0.91%
carbohydrate binding	1.37%	1.16%	0.00%	0.00%	6.36%	0.00%	9.80%	0.00%	0.76%	1.05%	2.66%
lipid binding	0.00%	5.81%	0.00%	0.00%	1.82%	11.76%	0.49%	6.25%	0.30%	0.00%	1.06%
chromatin binding	0.00%	0.00%	0.00%	0.00%	0.00%	0.00%	0.00%	0.00%	0.45%	1.05%	0.30%
transcription factor activity	8.22%	**20.93%**	0.00%	8.00%	6.36%	5.88%	1.96%	6.25%	4.38%	**22.11%**	6.84%
nuclease activity	2.74%	1.16%	0.00%	0.00%	0.00%	0.00%	0.00%	0.00%	0.15%	1.05%	0.38%
RNA binding	1.37%	1.16%	9.09%	4.00%	0.00%	0.00%	0.00%	0.00%	2.87%	2.11%	1.90%
translation factor activity, nucleic acid binding	0.00%	2.33%	0.00%	0.00%	0.00%	0.00%	0.00%	0.00%	1.21%	1.05%	0.84%
nucleotide binding	16.44%	5.81%	0.00%	12.00%	9.09%	17.65%	20.10%	12.50%	13.29%	17.89%	13.92%
oxygen binding	2.74%	2.33%	9.09%	8.00%	1.82%	5.88%	1.96%	3.13%	0.60%	1.05%	1.52%
protein binding	9.59%	11.63%	0.00%	8.00%	10.00%	17.65%	19.12%	9.38%	12.69%	11.58%	12.93%
motor activity	0.00%	1.16%	0.00%	0.00%	0.00%	0.00%	0.00%	0.00%	0.91%	2.11%	0.68%
receptor binding	0.00%	0.00%	9.09%	0.00%	0.00%	0.00%	0.00%	0.00%	0.00%	0.00%	0.08%
receptor activity	2.74%	0.00%	0.00%	0.00%	4.55%	0.00%	9.80%	6.25%	0.45%	1.05%	2.51%
structural molecule activity	0.00%	1.16%	9.09%	0.00%	0.00%	0.00%	0.00%	0.00%	13.29%	4.21%	7.15%
transcription regulator activity	1.37%	2.33%	0.00%	4.00%	0.00%	0.00%	0.49%	3.13%	0.76%	6.32%	1.29%
transporter activity	2.74%	8.14%	9.09%	12.00%	13.64%	11.76%	2.45%	3.13%	7.10%	3.16%	6.54%

first step of signal transduction in rice. In the molecular function, most tissue (organ)-specific genes are related to hydrolase, and over 30% plumule- and seedling- specific genes are involved in transferase activity (Fig. 2C). Interestingly, genes associated with TF activity were enriched in a developing panicle, reflecting the importance of transcription in panicle development. Wang *et al.* also made the same suggestion [3].

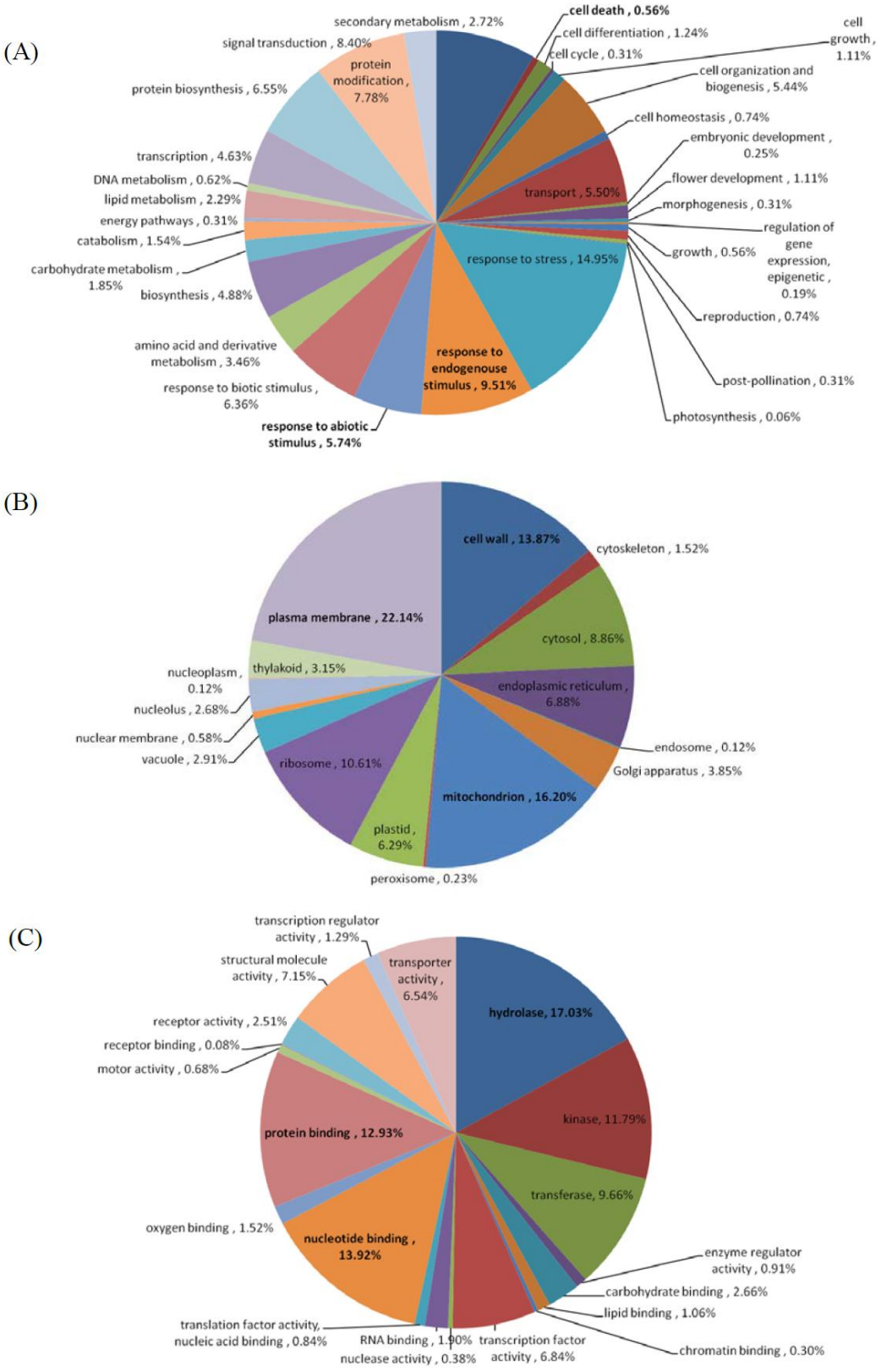


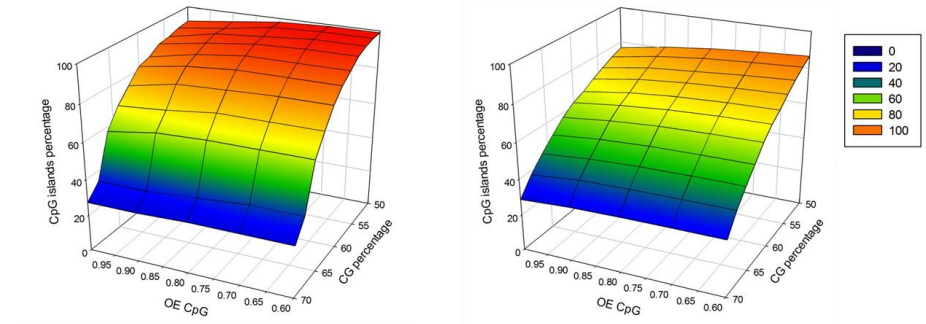
Figure 2 Functional annotations of tissue-specific genes with significantly over-represented GO terms.
(A) Biological Process (B) Cellular Component (C) Molecular Function

3.2. CpG/CpNpG islands and analysis of base composition

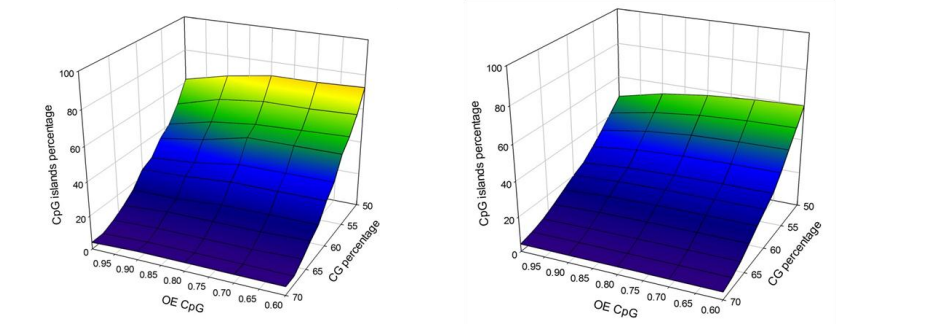
In the promoters of mammals, most tissue-specific genes lack CpG islands, and CpG islands are found mainly in least tissue-specific genes [16]. Based on these findings, this study attempts to determine whether CpG/CpNpG island is a critical feature for distinguishing tissue-specific from non-tissue-specific gene promoters in rice. Similarly, the percentage of CpG/CpNpG islands in non-tissue-specific gene promoters is more than tissue (organ)-specific one in rice when using various parameters to identify CpG/CpNpG islands (Table 3 and Fig. 3). However, several tissues (organs) have more than 50% promoters with CpG/CpNpG islands (based on a window size of 200 bp, OE CpG of 0.8, CG percentage of 60), especially for stamen, panicle, endosperm, and plumule with 74%, 73%, 67%, and 65%, respectively. It suggests that most genes in different developmental stages may require complicated regulation such as epigenetic modification in plant. Interestingly, most CpG/CpNpG islands are located near TSSs in both tissue (organ)-specific and ubiquitous genes (data not shown). Of particular interest in this study is whether the location of CpG/CpNpG islands is important to separating tissue (organ) and non-tissue (organ) specific promoters. Significantly, the percentage of CpG/CpNpG islands in ubiquitous genes exceeds 80%, the percentage of most tissue (organ)-specific genes is less than 50% according to the statistics of -200 bp to +200 bp and +1 bp to +200 bp regions (Fig. 4).

Table 3 The statistics of CpG/CpNpG islands in rice promoters

	Window size: 200 bp (OE CpG=0.8)		Window size: 500 bp (OE CpG=0.8)	
CG percentage	50	60	50	60
germination seed	74%	54%	49%	22%
endosperm	80%	67%	57%	30%
flower	72%	56%	42%	19%
leaf	68%	43%	34%	7%
panicle	85%	73%	66%	27%
plumule	88%	65%	53%	41%
root	75%	52%	38%	8%
seedling	85%	56%	41%	15%
stamen	89%	74%	64%	24%
stem	64%	44%	40%	12%
non-tissue(organ) specific	98%	88%	72%	27%



(A) Ubiquitous genes (Window size: 200 bp) (B) Tissue-specific genes (Window size: 200 bp)



(C) Ubiquitous genes (Window size: 500 bp) (D) Tissue-specific genes (Window size: 500 bp)

Figure 3 The percentage of CpG/CpNpG islands in ubiquitous (non-tissue specufic) and tissue- specific genes.(A) ubiquitous genes (based on window size = 200 bp) (B) Tissue-specific genes (based on window size = 200 bp) (C) ubiquitous genes (based on window size = 500 bp) (D) Tissue-specific genes (based on window size = 500 bp)

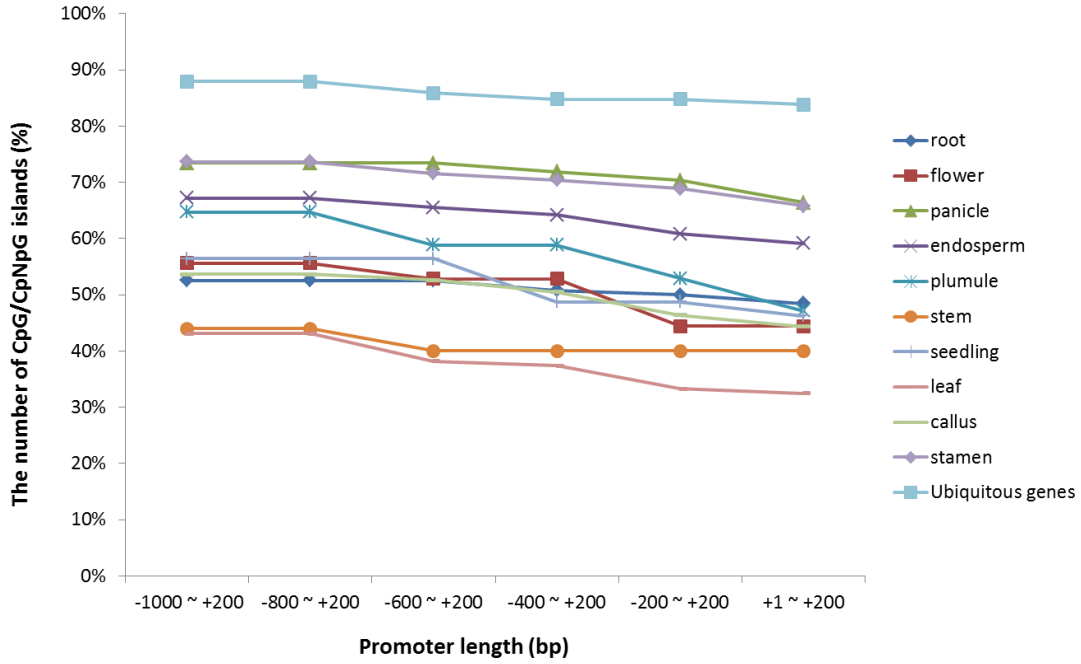


Figure 4 The number of CpG/CpNpG islands identified in different promoter regions in various tissues (organs) and ubiquitous (non-tissue specific) genes

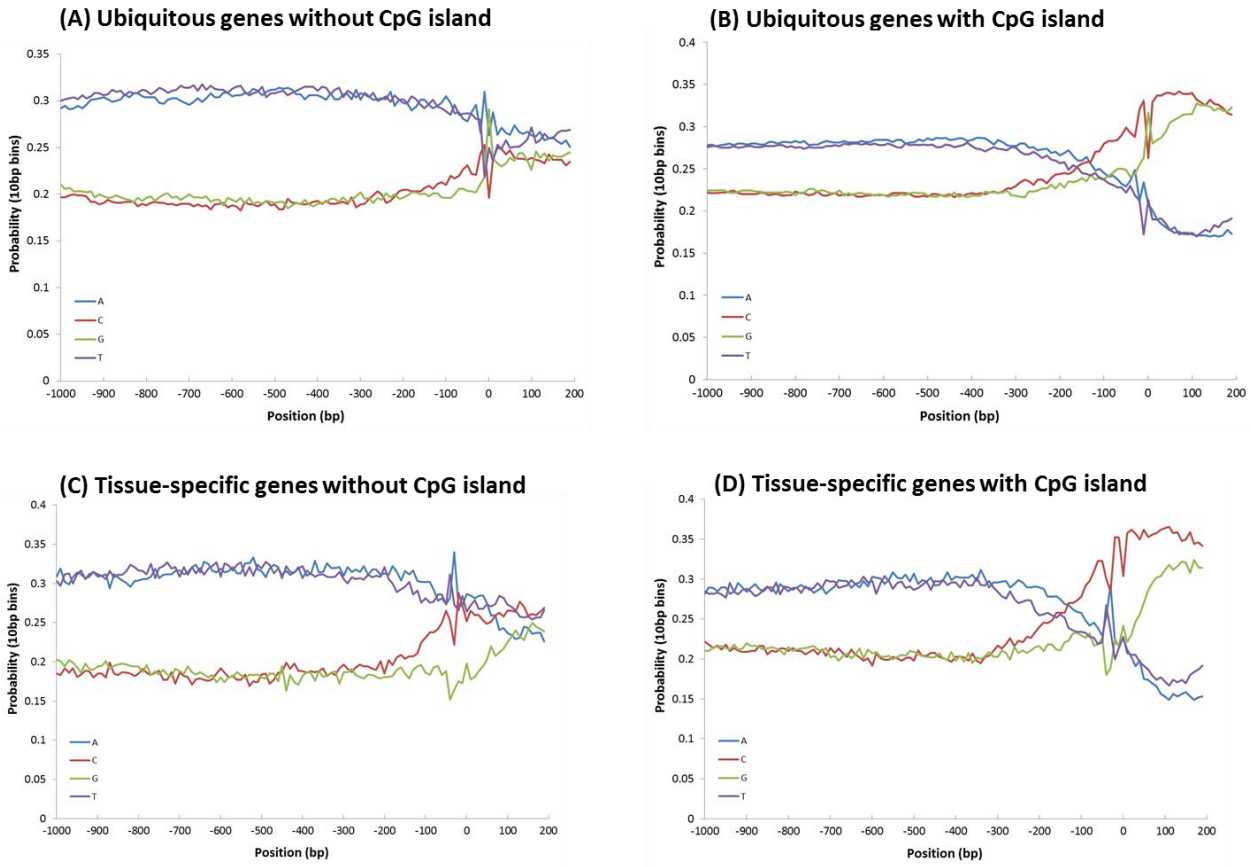


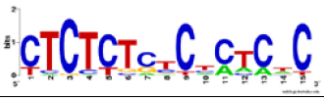
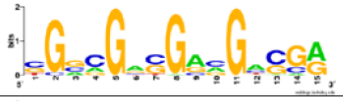
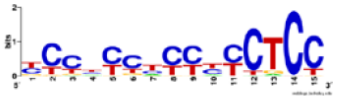
Figure 5 Base-composition profiles for ubiquitous and tissue- specific genes with and without CpG/CpNpG islands.(A) Ubiquitous genes without CpG/CpNpG island (B) Ubiquitous genes with CpG/CpNpG island (C) Tissue-specific genes without CpG/CpNpG island(D) Tissue-specific genes with CpG/CpNpGisland.

Furthermore, Schug*et al.* indicated that base-composition profiles of promoters help to distinguish tissue-specific from non-tissue-specific promoters in humans and mice [16]. This study also analyzed a full set of rice promoters to elucidate the base composition of non-specific and tissue-specific gene promoters. Figure 5 presents base composition profiles with window size of 10 bp. Notably, ubiquitous and tissue-specific gene promoters (with or without CpG/CpNpG islands) differ significantly in the base composition profiles in conjunction with CpG/CpNpG islands. The four classes of promoter considered below are ubiquitous-CGI+, ubiquitous-CGI-, tissue-specific-CGI+, and tissue-specific-CGI-. Although all four promoter classes have an A+T bias from -1000 bp to -100 bp, CGI+ and CGI- promoters significantly differ in A+T content (Fig. 5). Additionally, despite the A+T bias reaching almost p(A+T) = 0.65 from far up stream to -100 bp in the CGI- promoter, the bias is significantly lower in the CGI+ promoter, for which p(A+T) = 0.55. Conversely, the C+G bias increases substantially from -100 bp to +200 bp in the CGI+ promoter but not in the CGI- promoter (Fig. 5B and 5D). Moreover, a preference for C over G (p(C) >p(G)) exists in the downstream region of the tissue-specific-CGI+ promoter, while p(C) = p(G) in the ubiquitous-CGI+ promoter. However, the tissue-specific-CGI- promoter exhibits no obvious C+G bias, and p(C) >p(G) also in the +1 to +200 region (Fig. 5C). Additionally, the content of C is less than other nucleotides in the +1 to +200 region in ubiquitous-CGI- promoter that is converse in other class of promoter (Fig. 5A). Above differences between the base compositions of tissue (organ)-specific and nonspecific promoters in rice suggest that these promoters have different structural features and regulatory mechanisms.

3.3. Identification of tissue (organ)-specific motifs

As is well known, genes with a common biochemical function are associated with a set of gene promoters with some over-represented functional regulatory motifs [20]. Numerous motifs can be identified in tissue (organ)-specific promoters in this study (data not shown). Following removal of the redundancies by using TOMTOM program, motifs with a MEME E-value lower than 0.001 were selected as tissue (organ)-specific motifs and displayed in Table 4. Table 4 lists several tissue-specific motifs that were identified in various tissues. This table also displays numerous previously unknown motifs in the endosperm, leaf, panicle, and stamen. Since scientists have strangely neglected tissue (organ)-specific motifs in rice promoters, to our knowledge, no known rice TFBSs have been identified as tissue (organ)-specific motifs based on the straight criteria. In contrast, CDC5 and BZR1 in *Arabidopsis* were identified (Table 4). CDC5, a myb-related protein is critically involved in the cell cycles in yeast and animals. The protein is associated with spliceosome, and a previous study has posited that it has multiple regulatory functions, especially in the developmental and tissue-specific control of alternative splicing [21]. However, only a few attempts have so far been made to investigate CDC5 functions with plants [22]. Lin *et al*. indicated that although CDC5 has functions similar to yeast and human in *Arabidopsis* [22], no CDC5 ortholog of rice has been studied. This preliminary study attempts to determine whether CDC5 orthologs in rice have unique roles in tissue-specific control during development. Additionally, BZR1 is a tissue-specific expression protein in elongating cells [23]. Elongation cells are crucial for panicle elongation and development. Sunohara*et al.* demonstrated that panicle mutations significantly influence the development of culms and internodes in rice [24]. Therefore, the BZR1 motif can be regarded as a significant candidate panicle-specific motif in rice promoters, which is an interesting area for future research. Additionally, CxCCxCCxCC and GATxGAT motifs are highly conserved in stamen-specific promoters in rice. Both motifs are suggested to exam their function in stamen development, which is possibly applicable to male sterility research in agriculture. The motifs discovered in this study may be candidate TFBSs and have important roles in gene regulation during the various stages of development of rice. We are also going to design experiments (wet lab) to further confirm whether those unknown motifs are essential for tissue specific expression in the future.

Table 4 The motif logos of the tissue-specific structure present in different tissues

Tissues/Organs	Motif logo	Consensus sequence	Comment (E-value)
endosperm		CG[CT][CG]G[CG][CG][GTA]C[GC]C[ACG][GC][CG][GC]	CDC5, (2.80E-11), *Arabidopsis*
endosperm		CTTTTCCA[TC]CACATC	Unknown, (2.20E-09)
endosperm		GC[GCA]GC[GC]A[CT]G[GA][CT][GA][CAG][GC][GC]	Unknown, (4.60E-09)
endosperm		[AG][AG][AG]A[CT]GGAGG[GT]AGTA	Unknown, (1.10E-04)
leaf		[CT]TCTCT[CG][TC]C[TCA][CA][TA]C[TA]C	Unknown, (5.40E-11)
panicle		[CG]G[GCA][CA]G[AG][CG]G[AG][CAG]G[AC][CG][GC][AG]	Unknown, (6.00E-16)
panicle		C[GA]AAT[GA]TTTG[GA]ACAC	Unknown, (4.90E-04)
panicle		GTTTG[GA][AG]AA[GA]C[GA]TGC	BZR1, (5.60E-04) *Arabidopsis*
stamen		[TC][CT][CT][TC][CT][CT][TC][CT][CT][TC][CT]CTCC	Alfin1, (1.3e-323), *alfalfa*
stamen		[CG]C[TGA]CC[TA]CC[TAG]CC	Unknown, (1.30E-83)
stamen		ATGTTTACTGTAGCA	Unknown, (2.20E-70)
stamen		[CG]GATCGAT[CG]GA	Unknown, (3.70E-62)
stamen		AAA[CG][AT]TT[TC]GATGTGA	Unknown, (8.70E-62)

4. ACKNOWLEDGMENTS

The authors sincerely appreciate the National Science Council of the Republic of China, for financially supporting this research under Contract Numbers of NSC 99-2621-B-006 -001 -MY2 and NSC 99-2628-B-006 -016 -MY3. Ted Knoy is appreciated for his editorial assistance.

5. REFERENCES

[1] V. A. Benedito, I. Torres-Jerez, J. D. Murray, A. Andriankaja, S. Allen, K. Kakar, M. Wandrey, J. Verdier, H. Zuber, T. Ott, S. Moreau, A. Niebel, T. Frickey, G. Weiller, J. He, X. Dai, P. X. Zhao, Y. Tang, and M. K. Udvardi, "A gene expression atlas of the model legume Medicago truncatula," *Plant J*, vol. 55, pp. 504-13, Aug 2008.

[2] M. Schmid, T. S. Davison, S. R. Henz, U. J. Pape, M. Demar, M. Vingron, B. Scholkopf, D. Weigel, and J. U. Lohmann, "A gene expression map of Arabidopsis thaliana development," *Nat Genet*, vol. 37, pp. 501-6, May 2005.

[3] L. Wang, W. Xie, Y. Chen, W. Tang, J. Yang, R. Ye, L. Liu, Y. Lin, C. Xu, J. Xiao, and Q. Zhang, "A dynamic gene expression atlas covering the entire life cycle of rice," *Plant J*, vol. 61, pp. 752-66, Mar 2010.

[4] 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 TRANSCompel: transcriptional gene regulation in eukaryotes," *Nucleic Acids Res*, vol. 34, pp. D108-10, Jan 1 2006.

[5] D. Takai and P. A. Jones, "The CpG island searcher: a new WWW resource," *In Silico Biol*, vol. 3, pp. 235-40, 2003.

[6] W. C. Chang, T. Y. Lee, H. D. Huang, H. Y. Huang, and R. L. Pan, "PlantPAN: Plant promoter analysis navigator, for identifying combinatorial cis-regulatory elements with distance constraint in plant gene groups," *BMC Genomics*, vol. 9, p. 561, 2008.

[7] T. Obayashi, K. Kinoshita, K. Nakai, M. Shibaoka, S. Hayashi, M. Saeki, D. Shibata, K. Saito, and H. Ohta, "ATTED-II: a database of co-expressed genes and cis elements for identifying co-regulated gene groups in Arabidopsis," *Nucleic Acids Res*, vol. 35, pp. D863-9, Jan 2007.

[8] Y. A. Chen, Y. C. Wen, and W. C. Chang, "AtPAN: an integrated system for reconstructing transcriptional regulatory networks in Arabidopsis thaliana," *BMC Genomics*, vol. 13, p. 85, 2012.

[9] G. Haberer, M. T. Mader, P. Kosarev, M. Spannagl, L. Yang, and K. F. Mayer, "Large-scale cis-element detection by analysis of correlated expression and sequence conservation between Arabidopsis and Brassica oleracea," *Plant Physiol*, vol. 142, pp. 1589-602, Dec 2006.

[10] D. Walther, R. Brunnemann, and J. Selbig, "The regulatory code for transcriptional response diversity and its relation to genome structural properties in A. thaliana," *PLoS Genet*, vol. 3, p. e11, Feb 9 2007.

[11] L. Gautier, L. Cope, B. M. Bolstad, and R. A. Irizarry, "affy--analysis of Affymetrix GeneChip data at the probe level," *Bioinformatics*, vol. 20, pp. 307-15, Feb 12 2004.

[12] A. Vandenbon and K. Nakai, "Modeling tissue-specific structural patterns in human and mouse promoters," *Nucleic Acids Res*, vol. 38, pp. 17-25, Jan 2010.

[13] S. Laubinger, G. Zeller, S. R. Henz, T. Sachsenberg, C. K. Widmer, N. Naouar, M. Vuylsteke, B. Scholkopf, G. Ratsch, and D. Weigel, "At-TAX: a whole genome tiling array resource for developmental expression analysis and transcript identification in Arabidopsis thaliana," *Genome Biol*, vol. 9, p. R112, 2008.

[14] S. Ouyang, W. Zhu, J. Hamilton, H. Lin, M. Campbell, K. Childs, F. Thibaud-Nissen, R. L. Malek, Y. Lee, L. Zheng, J. Orvis, B. Haas, J. Wortman, and C. R. Buell, "The TIGR Rice Genome Annotation Resource: improvements and new features," *Nucleic Acids Res*, vol. 35, pp. D883-7, Jan 2007.

[15] 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*, vol. 25, pp. 25-9, May 2000.

[16] J. Schug, W. P. Schuller, C. Kappen, J. M. Salbaum, M. Bucan, and C. J. Stoeckert, Jr., "Promoter features related to tissue specificity as measured by Shannon entropy," *Genome Biol*, vol. 6, p. R33, 2005.

[17] L. B. Timothy and E. Charles, "Fitting a mixture model by expectation maximization to discover motifs in biopolymers," in *Proceedings of the Second International Conference on Intelligent Systems for Molecular Biology*, ed. Menlo Park, California: AAAI Press, 1994, pp. 28-36.

[18] S. Gupta, J. A. Stamatoyannopoulos, T. L. Bailey, and W. S. Noble, "Quantifying similarity between motifs," *Genome Biol*, vol. 8, p. R24, 2007.

[19] G. E. Crooks, G. Hon, J. M. Chandonia, and S. E. Brenner, "WebLogo: a sequence logo generator," *Genome Res*, vol. 14, pp. 1188-90, Jun 2004.

[20] G. Meng, A. Mosig, and M. Vingron, "A computational evaluation of over-representation of regulatory motifs in the promoter regions of differentially expressed genes," *BMC Bioinformatics*, vol. 11, p. 267, 2010.

[21] X. H. Lei, X. Shen, X. Q. Xu, and H. S. Bernstein, "Human Cdc5, a regulator of mitotic entry, can act as a site-specific DNA binding protein," *J Cell Sci*, vol. 113 Pt 24, pp. 4523-31, Dec 2000.

[22] Z. Lin, K. Yin, X. Wang, M. Liu, Z. Chen, H. Gu, and L. J. Qu, "Virus induced gene silencing of AtCDC5 results in accelerated cell death in Arabidopsis leaves," *Plant Physiol Biochem*, vol. 45, pp. 87-94, Jan 2007.

[23] Z. Y. Wang, T. Nakano, J. Gendron, J. He, M. Chen, D. Vafeados, Y. Yang, S. Fujioka, S. Yoshida, T. Asami, and J. Chory, "Nuclear-localized BZR1 mediates brassinosteroid-induced growth and feedback suppression of brassinosteroid biosynthesis," *Dev Cell*, vol. 2, pp. 505-13, Apr 2002.

[24] H. Sunohara, H. Satoh, and Y. Nagato, "Mutations in panicle development affect culm elongation in rice," *Breeding Science*, vol. 53, pp. 109-117, Jun 2003.

TWO SEGMENTATION METHODS FOR GENOME ANNOTATION

Alice Cleynen[1,2]*, Sandrine Dudoit*[3]*, Emilie Lebarbier*[1,2] *and Stéphane Robin*[1,2]

[1]Agroparistech, UMR 518, 16 rue Claude Bernard, 75005 Paris, France,
[2]Inra, UMR 518, 16 rue Claude Bernard, 75005 Paris, France,
[3]Division of Biostatistics and Department of Statistics,
University of California, Berkeley, 185 Li Ka Shing Center, #3370, Berkeley, CA 94720-3370, USA
alice.cleynen@agroparistech.fr, sandrine@stat.berkeley.edu,
emilie.lebarbier@agroparistech.fr, robin@agroparistech.fr

ABSTRACT

We present two segmentation methods using the negative binomial distribution to address two biological questions related to sequencing data: the assessment of the number and localization of expressed genes based on RNA-Seq data, and the precise and confident gene re-annotation. Our first algorithm computes a penalized log-likelihood estimator of the regression function with a complexity of $\mathcal{O}(Kn\log n)$ with theoretical bounds for its efficiency. Our second algorithm computes in a Bayesian framework the exact posterior probabilities of the change-point location with quadratic complexity. We illustrate the results of those methods on simulation studies inspired by RNA-Seq datasets. Both methods are available as R packages on the CRAN repository.

1. INTRODUCTION

Our motivating example is the analysis of RNA-Seq data. The output of an RNA-Seq experiment is the number of reads (i.e. short portions of the genome) which first position maps to each location of a genome of reference. Supposing that we dispose of such a sequence, we expect to observe a stationarity in the amount of reads falling in different areas of the genome: coding sequences, intronic regions, etc. We wish to localize those regions that are biologically significant. This problem can be seen as a multiple change-point detection setting for count datasets, and can be written as follows: we observe a finite sequence of size n, $\{y_t\}_{t\in\{1,\dots,n\}}$ realization of independent variables Y_t which are supposed to be drawn from the negative binomial distribution $\mathcal{NB}$, adapted to the RNA-seq experiment analysis [1]:

$$Y_t \sim \mathcal{NB}(p_t, \phi),\ 1 \leq t \leq n,$$

where the parameters $\{p_t\}$ are assumed to be piece-wise constant and so subject to an unknown number $K-1$ of abrupt changes occurring at change-point locations $\{\tau_k\}$, and ϕ is a constant parameter corresponding to the dispersion in the sequence. Thus, there is a partition of $\{1,\dots,n\}$ into K segments within which the observations follow the same distribution and between which observations have different distributions, i.e. p_t is constant within a segment J with value p_J and differs from a segment to another.

We consider two biological questions. The first, (i), is inspired from whole-genome analysis where we are interested in the localization of transcribed regions on chromosomes with signal length ranging from 10^6 to 10^8 datapoints. In this situation, the number of segments is typically unknown (for instance the number of genes transcribed, or the number of exons of the expressed isoforms is not known) and might be a question in itself. The second, (ii), is inspired from transcript re-annotation where we want to precisely and confidently localize the exon / intron boundaries from a signal surrounding a gene with length of the order of 10^4. In this scenario, we can assume the number of segments to be known from previous annotation. We propose two segmentation methods to address these two biological questions. The first, PDPA, is a penalized log-likelihood estimator for negative binomial distributed datasets which satisfies an oracle inequality in a non-asymptotic context and which algorithmic complexity allows its use in our first framework (i). The second, EBS, is a Bayesian segmentation method providing the exact computation of posterior probabilities of change-point location at the price of higher complexity, thus restricted to our second framework (ii).

2. METHODS

In the next sections we will denote by m a partition of $[\![1,n]\!]$, $\quad m = \{[\![1,\tau_1[\![, [\![\tau_1,\tau_2[\![, \dots, [\![\tau_{K-1},n]\!]\} = \{1,\tau_1,\dots,\tau_{K-1},n+1\}$ with $|m| = K$ segments, by J a segment of m and $\mathcal{M}_K$ will be the set of all possible partitions of $[\![1,n]\!]$ in K segments.

2.1. Whole-genome analysis

In framework (i) we want to estimate the distribution s of the data, $s(t) = \mathcal{NB}(p_t,\phi)$ through a segmentation $\hat{m}$ such that

$$\forall J \in \hat{m}, \forall t \in J,\ \hat{s}(t) = \mathcal{NB}(p_J,\phi). \tag{1}$$

The main difficulty is the choice of the segmentation $\hat{m}$, since the parameters $\{p_J\}$ can often be estimated trivially

by maximum likelihood given estimates of the $\{\tau_k\}$s. Following [2] and noting $\gamma(u)$ the log-likelihood of distribution u, we propose to choose $\hat{m}$ as

$$\hat{m} = \arg\min_{m \in \mathcal{M}} \{\gamma(\hat{s}_m) + pen(m)\},$$

for 'good' choices of model collection $\mathcal{M}$ and of penalty function pen, i.e. such that the resulting estimator satisfies an oracle inequality. More specifically, we can hope that up to a constant, it performs as well in terms of Hellinger risk as the best but unreachable estimator.

For any given K, and with empirical complexity of $\mathcal{O}(Kn \log n)$, the Pruned Dynamic Programming Algorithm (PDPA,[3]) computes the best estimator $\hat{s}_K$ among the collection $\mathcal{S}_K = \bigcup_{m \in \mathcal{M}_K} \{s_m \mid \forall J \in m,\ \forall t \in J,\ s_m(t) = \mathcal{G}(\theta_J)\}$ with respect to γ for one-parameter unimodal distributions $\mathcal{G}$ from the exponential family. The negative binomial distribution is included in the latter on the condition that the over-dispersion parameter ϕ is known. We therefore propose to use a modified Jonhson and Kotz's estimator [4] for ϕ. Specifically, for each sliding window of size h equal to twice the size of the longest zero band, we compute the method of moments estimator of ϕ, using the formula $\phi = \mathbf{E}^2(X)/(\mathbf{V}(X) - \mathbf{E}(X))$, and retain the median over all windows. We implemented the PDPA in an R package [5] for distributions including the negative binomial for which we included our estimator.

Among the so-constructed collection of estimators $\mathcal{S} = \{\hat{s}_K\}_{1 \leq K \leq K_{max}}$, we have to choose the best one, i.e. choose an optimal number of segments K via a good choice of the penalty function. To this end, we show [6] that for

$$pen(m) = \beta |m| \left(1 + 4\sqrt{1.1 + \log\left(\frac{n}{|m|}\right)}\right)^2, \qquad (2)$$

with β a constant to be tuned according to the data, and up to a $\log n$ factor, we satisfy an oracle inequality.

As the penalty we propose depends on the segmentation only through its size, our final algorithm is two-steps: first we estimate for $1 \leq K \leq K_{max}$, the location of the change-points $\{\tau_k\}$ and the parameters $\{p_J\}$ and ϕ using the PDPA, then we estimate the number of segments K using our penalty function and the slope heuristic [7] to tune β. This procedure is automated in our R-package. We illustrate the performances of our algorithm on simulation studies in the Result Section.

2.2. Gene re-annotation

One can be interested in the confidence of the proposed estimator, for instance in the context of gene re-annotation (ii). At the price of higher complexity (quadratic), [8] proposed a Bayesian segmentation method which, among other quantities of interest, computes the exact posterior probabilities of the k^{th} change-point occurring at each location t given a number of segments K. It relies on operations on the triangular matrix A which generic term is, for $i < j$, $[A]_{i,j} = P(Y_{[\![i,j[\![} | [\![i, j[\![)$, and requires the ability to compute $[A]_{i,j}$ exactly.

We implemented this algorithm in an R package, `EBS`, for distributions including the negative binomial, overcoming two major difficulties: the numerical precision required to deal with extremely small probabilities due to the length of the signals (up to 10^4 data-points), and the assumptions required for the exact computation of posterior probabilities (knowledge and thus estimation of the overdispersion parameter ϕ using the estimator described in the previous paragraph, and existence of a conjugate prior for the parameters to segment).

Given a segmentation m, our model can be written as follows:

$$\begin{aligned} \forall J \in m, \qquad & p_J \sim \mathcal{B}eta(a, b) \\ \forall J \in m, \forall t \in J, \qquad & Y_t \sim \mathcal{NB}(p_J, \phi) \end{aligned}$$

With convenient choice of hyper-parameters (such as Jeffrey's prior $p_J \sim \mathcal{B}eta(1/2, 1/2)$) we can compute, for all $1 \leq t \leq n$, $p(\tau_k = t | Y, K)$ and the associated 95% credibility intervals. We illustrate our results on simulation studies presented in the next section.

3. RESULTS

3.1. Whole-genome analysis

In framework (i), we assess the quality of our method PDPA on two simulated datasets. Using the SGD annotation (`http://www.yeastgenome.org/`), we have constructed a signal of length 230218 with 119 segments corresponding to the size of the yeast positive strand of chromosome 1 and its 59 annotated genes. In a first scenario we have simulated the datasets from the negative binomial distribution. After running our overdispersion estimator on each strand of the 16 sequenced chromosomes of our yeast dataset and keeping its median value, we set ϕ to 0.27 and we randomly chose the values of the parameter p_J on each segment between four possibilities to mimic a real dataset: 0.9 for non-coding regions, and respectively 0.25, 0.1 and 0.05 for low, medium and highly expressed transcripts. In a second scenario we simulated the datasets by resampling at random and with replacement from four groups of real RNA-Seq data previously pooled into classes of expression: intronic, low, medium and high.

PDPA selected respectively 117 and 118 segments, with a runtime of about 25 minutes on a standard laptop. Figure 1 displays the percentage of recovered true change-points as the number of segment increases, as well as the choice of $\hat{K}$ given by our penalty function. In both cases, the choice corresponds to the largest possible K before adding segments does not increase the percentage of true positives. This indicates both the stability of the log-likelihood criterion and the pertinence of the penalty function.

Running our algorithm on the real data-set leads to a selection of 103 segments, most of which surrounds known genes from the SGD annotation. Only three change-points were classified as false positives. Figure 2 illustrates the result.

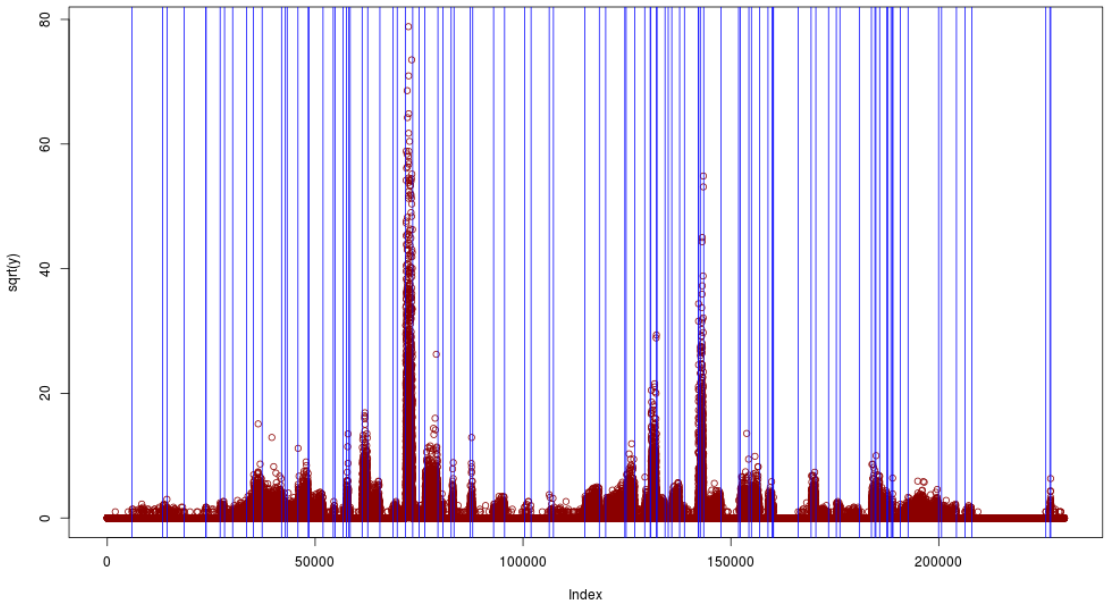


Figure 2. *Proposed segmentation of a real data-set.* Example of our proposed segmentation for the chromosome 1 of the yeast organism. Read counts are plotted with a squared-root scale, blue lines indicate estimated change-points.

3.2. Gene re-annotation

We illustrate our second method EBS on two simulation studies using our RNA-Seq data. First we resampled (with replacement) each region of the signal surrounding the yeast gene YAL035W according to the SGD annotation, to obtain a segmentation in 5 segments (the top left figure of Figure 3 illustrates the true data for this gene and the posterior distribution of the change-points (in blue) obtained with the EBS algorithm.) In our second scenario, we created an artificial gene, inspired from the *Drosophila Melanogaster* inr-a gene, resulting in a 14-segment signal with irregular intensities mimicking a differentially transcribed gene (the bottom left figure illustrate such a simulated data-set, and the output of EBS). Each configuration was simulated 100 times, with processing run-time averaging at respectively 7 seconds and 20 minutes.

We evaluated the false positive and false negative rates by declaring t a break-point if $\exists 1 \leq k < K$, $p(\tau_k = t|Y, K) \geq \lambda$ for a given threshold λ and by varying λ and averaging the resulting proportions of false positives and false negatives over simulations. A perfect ROC curve should indicate a sharp change-point posterior probability located at the exact expected position. In our case it leads to the ROC-like curves as presented in the right side of Figure 3. The latter have an almost perfect shape confirmed by the average credibility interval sizes and proportion of times that each covers its associated true change-point. Examples of those values are given in Table 1.

Gene	Interval length	Coverage
YAL035W	10	0.97
Inr-a	7	0.99

Table 1. *Credibility intervals.* Median length of the 95% credibility intervals and percentage of simulations for which the intervals covered the true first break-point (out of 100).

3.3. Comparison to existing methods

In a recent analysis (paper under review), we compared our two methods with other approaches adapted to count datasets: CART [9, 10], a fast heuristic algorithm and PELT [11], an exact algorithm where the number of segments is estimated within the algorithm both implemented for the Poisson distribution, and postCP [12], a constrained hidden Markov model (HMM) approach for segmentation which uses the PDPA for its parameter initialization. We showed that PDPA outperforms other methods, especially when the number of segments is known, both on simulated and real-data. At the cost of higher complexity, EBS had excellent results on ROC-like curves, allowing its use for posterior applications such as transcript location comparisons.

4. CONCLUSION

We have presented two segmentation approaches using the negative binomial distribution to address biological questions related to the analysis of sequencing data. Our first method, PDPA, allows to assess the number and location of transcribed regions in an RNA-Seq experiment on lower organisms such as yeast. Our resulting estimator not only satisfies theoretical oracle inequalities but in practice selects a number of segments and an associated segmen-

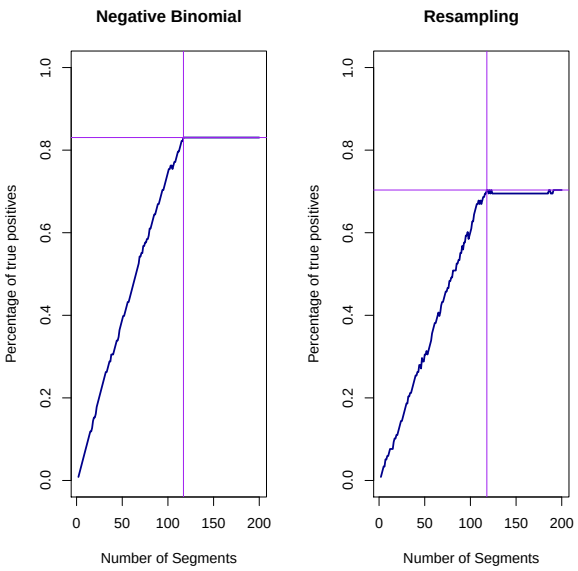


Figure 1. *Estimation of s on long simulations.* For both simulation study (Left: simulation from negative binomial distribution, Right: simulation by resampling real RNA-Seq data), percentage of true change-points recovered by the segmentation as the number of segments increases. Purple lines indicate the number of segments chosen by the penalty function and the corresponding percentage of true change-points.

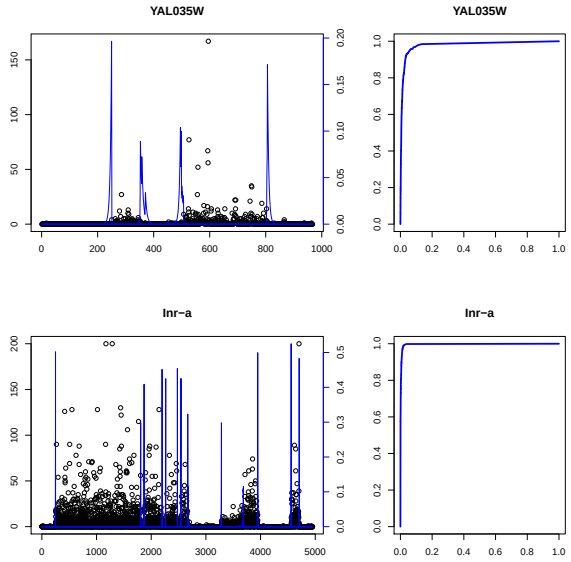


Figure 3. *Posterior change-point probabilities.* Left: Example of posterior change-point location probabilities for gene YAL035W from the yeast (Top) and Inr-a from the Drosophila (Bottom). Right: ROC-like curves obtained by averaging over the 100 simulations from YAL035W (Top) and Inr-a (Bottom).

tation which is optimal according to the data and with a complexity allowing its use on long signals. While these results are likely to be compromised in higher organisms where genes are subject to extensive alternative splicing and exons from different genes can overlap, we can hope that in such context, the use of existing annotation or other algorithms like Cufflinks will overcome these difficulties.

Our second algorithm, EBS, allows to address the quality and confidence of the proposed estimator with excellent results in terms of ROC-curves and coverage of credibility intervals. With higher complexity, it remains usable for the re-annotation of genes. Both algorithms are available as R packages with full documentation on the CRAN repository.

5. AVAILABILITY OF SUPPORTING DATA

The dataset supporting the results of this article is available in the Sequence Read Archive repository, `http://www.ncbi.nlm.nih.gov/sra`, with the accession number SRA048710.

6. ACKNOWLEDGEMENTS

The authors wish to thank M. Koskas and G. Rigaill for their help with writing the code and for numerous helpful conversations. They would also like to thank the anonymous reviewers for their valuable comments and suggestions to improve the quality of the paper.

Alice Cleynen's research was supported by an Allocation Special Normalien at the Université Paris-Sud in Orsay.

7. REFERENCES

[1] D. Risso, K. Schwartz, G. Sherlock, and S. Dudoit, "GC-content normalization for RNA-Seq data," *BMC Bioinformatics*, vol. 12, no. 1, pp. 480, 2011.

[2] L. Birgé and P. Massart, "Gaussian model selection," *J. Eur. Math. Soc. (JEMS)*, vol. 3, no. 3, pp. 203–268, 2001.

[3] G. Rigaill, "Pruned dynamic programming for optimal multiple change-point detection," *Arxiv preprint arXiv:1004.0887*, under review.

[4] N. Johnson, A. Kemp, and S. Kotz, "Univariate discrete distributions," *John Wiley & Sons, Inc.*, 2005.

[5] A. Cleynen, M. Koskas, and G. Rigaill, "A generic implementation of the pruned dynamic programing algorithm," *Arxiv preprint arXiv:1204.5564*, under review.

[6] A. Cleynen and E. Lebarbier, "Segmentation of the poisson and negative binomial rate models: a penalized estimator," *Arxiv preprint arXiv:1301.2534*, under review.

[7] S. Arlot and P. Massart, "Data-driven calibration of penalties for least-squares regression," *J. Mach. Learn. Res.*, vol. 10, pp. 245–279 (electronic), 2009.

[8] G. Rigaill, E. Lebarbier, and S. Robin, "Exact posterior distributions and model selection criteria for multiple change-point detection problems," *Statistics and Computing*, vol. 22, pp. 917–929, 2012.

[9] A. Scott and M. Knott, "A cluster analysis method for grouping means in the analysis of variance," *Biometrics*, vol. 30, pp. 507–512, 1974.

[10] Breiman, Friedman, Olshen, and Stone, "Classification and regression trees.," *Wadsworth and Brooks*, 1984.

[11] R. Killick, P. Fearnhead, and I. Eckley, "Optimal detection of changepoints with a linear computational cost," *Journal of the American Statistical Association*, , no. just-accepted, 2012.

[12] T. M. Luong, Y. Rozenholc, and G. Nuel, "Fast estimation of posterior probabilities in change-point models through a constrained hidden Markov model," *Arxiv preprint arXiv:1203.4394*, under review.

SENSITIVITY ANALYSIS FOR AN ODE BASED SYSTEM MODELING T LYMPHOCYTES

Giovanni Dalmasso[1], *Juliane Mai*[1] *and Sabine Attinger*[1]

[1]Department Computational Hydrosystems, Helmholtz-Centre for Environmental Research - UFZ, Permoserstrasse 15, 04318 Leipzig, Germany
giovanni.dalmasso@ufz.de

ABSTRACT

Differentiation processes of the immune system's cells have been shown to be affected by several environmental factors. Yet the underlying mechanism behind the differentiation processes of the T lymphocyte cells remain elusive. Here we use a mathematical model based on ordinary differential equations in order to investigate the influence of parameters on such biological models. We perform a global sensitivity analysis to identify the parameters which have a major impact on the model. We find the feasible ranges of each parameter discovering those that influence the model most. The present findings underline the importance of a prior sensitivity analysis in order to increase the efficiency and reliability of parameter inference of complex systems.

1. INTRODUCTION

Environmental chemicals ingested or inhaled from different sources like food or drinking water, industrial-, automobile exhaust can have a major impact on our immune system [1].

A particularly important group of immune cells are so called T lymphocyte cells. Adaptive immune system is mainly governed by T cells which play an essential function in the protection against pathogens, cancer and autoimmune diseases, among others [2]. A complex network of different kind of lymphocytes underlies the regulation of the protection in the organism and an appropriate balance in the differentiation process of T cells is decisive for the global health in humans and animals [3].

To explain the underling mechanisms in the immune system, disparate models have been elaborate, either using a mathematical approach [4] or combining mathematics with experiments [5]. However, the majority of these studies are performed in mice and, due to the lack of samples, only a few are focused on humans [6].

So far, established models rely on a wide range of parameters [7]. Since the inversion of parameters of complex models can not be performed reliably in presence of only a few data-points, the amount of parameters has to be reduced. It is then essential in this context to identify the most relevant parameters in order to reduce the model complexity along with experimental effort required and, therefore, produce a more efficient inversion strategy. In this work, we focus our study on the investigation of parameters' impact aiming to reduce the model complexity. The model of Busse et al. [5] is used as a benchmark for testing our approach. Furthermore, to overcome the weakness of many common biological models regarding the numerical method adopted to solve the governing model equations [8], we implement an adaptive time-step solver for ordinary differential equations (ODEs) in order to obtain a predetermined accuracy with minimal computational power.

2. MATERIALS AND METHODS

Mathematical models in biology are often based on systems of coupled ODEs. So far, established models able to describe differentiation processes of cells belonging to the immune system, rely on a wide range of different parameters, most of those need to be determined experimentally. The number of these parameters is usually larger than the data points achievable from the measurements. Consequently, it is essential identifying the most relevant ones in order to reduce the model complexity by performing a prior global sensitivity analysis of the model of interest.

2.1. ODE system

In the most generic form, any initial-value problem of n evolution equations with n initial conditions u_{i0} ($i = 1, \ldots, n$), could be written as

$$\begin{cases} \dfrac{d\mathbf{u}}{dt} &= \mathbf{f}(t, \mathbf{u}), \\ \mathbf{u}(t_0) &= \mathbf{u}_0 \end{cases} \tag{1}$$

where

$$\mathbf{u}: \begin{array}{c} \mathbf{R}_+ \longrightarrow \mathbf{R}^n \\ t \longmapsto \mathbf{u}(t) \end{array} \quad \mathbf{u}(t) \in \mathcal{D}_u \subseteq \mathbf{R}^n,\ t \in \mathbf{R}_+ \,. \tag{2}$$

As test model for this work, we choose the reaction-diffusion system proposed in Busse et al. [5] for the number of unoccupied cytokine interleukin-2 receptors (IL-2R) R, the number of receptor complex with cytokine interleukin-2 (IL-2) C and the number internalized IL-2R

per cell E:

$$\begin{cases} \dfrac{dR}{dt} &= v - k_{iR}R - k_{on}RI \\ & \quad +k_{off}C + k_{rec}E \\ \dfrac{dC}{dt} &= k_{on}RI - (k_{iC} + k_{off})\,C \\ \dfrac{dE}{dt} &= k_{iC}C - (k_{rec} + k_{deg})\,E \end{cases} \tag{3a}$$

with, the initial conditions

$$\begin{cases} R(0) &= 0.1 \\ C(0) &= 0.5 \\ E(0) &= 0.5 \end{cases} \tag{3b}$$

where the IL-2R production v is defined by

$$v = v_0 + v_1 \frac{C^m}{K^m + C^m}\,. \tag{3c}$$

The system (3), in the work of Busse et al. [5], is used to model both T helper (Th) cells and regulatory T (Treg) cells. The analysis of Treg cells is omitted here since the main focus of our work is a deep analysis of the parameters involved in the system more than the behavior of these two different groups of cells. The parameters involved in (3) are summarized in Table 1.

Symbol	Parameter
$p_1 : k_{iR}$	Internalization rate constant of IL-2R
$p_2 : k_{on}$	IL-2 association rate constant to IL-2R
$p_3 : k_{off}$	IL-2 dissociation rate constant from IL-2R
$p_4 : k_{rec}$	Recycling rate constant of IL-2Rα
$p_5 : k_{ic}$	Internalization constant of IL-2/IL-2R complex
$p_6 : k_{deg}$	Endosomal degradation constant IL-2R
$p_7 : v_0$	IL-2 receptor dynamics
$p_8 : v_1$	Feedback induced IL-2 receptor expression rate
$p_9 : K$	Half-saturation constant
$p_{10} : m$	Hill coefficient
$p_{11} : I$	IL-2 concentration

Table 1. Parameters involved in the model proposed by Busse et al. [5] (Eq. (3)).

In particular, we confine our study on the number of unoccupied IL-2R R. In the system (3a), the term I represents a diffusion equation [5]. However, since in these models the diffusion occur usually almost instantaneously, it was already shown by Busse et al. [5] that it could be considered as a constant. Therefore, we neglect the diffusion equation replacing it with constant and incorporating it in the parameters' set.

Besides, to achieve an optimal accuracy in the solution of system (3) with minimum computational effort, we implemented a 4^{th}-order Runge-Kutta method with adaptive step-size control.

2.2. Feasible parameter ranges

Finding feasible ranges of the parameters involved in biological models is often demanding. Both data from experiments and from literature could be not enough to fill the lack of parameters needed to describe these systems.

In this section we introduce an approach adopted to overcome this issue, namely, analyzing which parameters influence the model most in order to increase the reliability of parameter inference of complex models.

We know a priori the shape of the solution of our system (Figure 1, 3) and search for values that lead to this profile in a temporal interval $t \in [0, 100]$. To have a rough estimation of the possible ranges, we first consider guessed intervals according to the values used in the work of Busse et al. [5]. From these intervals we randomly generate a set of uniformly distributed parameters using the Brent-xor4096 algorithm [9]. We then run our model and discard all the ones that generate not feasible curves. Namely, parameters that produce curves that do not satisfy a number of fixed constraints previously imposed, i. e. not reproducing the shape of the reference solution (Figure 1). Analyzing the tick curve in Figure 1 more in detail, it is possible to identify two main parts of the solution. In the first, the curve slowly increases until it reaches the time point around 50, then it hastily changes curvature and, in the second part, it progressively stabilizes and attains a plateau.

After this first evaluation of feasible ranges, we repeat the same analysis changing several parameters at the same time. In this way we can verify if modifying combinations of parameters affects the model and the ranges of feasible values differently. Hence, the estimated parameters ranges are independent from parameters considered separately and take into account parameters correlation. Therefore, the results are more reliable.

2.3. Sensitivity analysis: Elementary Effects

The *Elementary Effects* (EE) method, produces a qualitative sensitivity analysis evaluation, in particular, is able to identify non-significant inputs or to order input components in relation to their influence in the system under investigation. Due to these reasons, EE method is one of the most applied screening method in sensitivity analysis, especially if the model examined is composed of an extensive quantity of parameters.

Since the system (3) is highly dependent on the parameters setting, we use the EE method described in this section, to measure their influence on the model.

Any EE method is defined as:

$$EE_j = \frac{M(p_1, \ldots, p_j + \Delta, \ldots, p_N)}{\Delta} - \frac{M(p_1, \ldots, p_j, \ldots, p_N)}{\Delta} \tag{4}$$

where EE_j is the EE of the j^{th} parameter, M is the model dependent on the N parameters $p_1, \ldots, p_N$ and Δ is the change of the j^{th} parameter. Given N parameters and K single samples to evaluate the EE of a defined parameter j,

$2NK$ parameter sets are needed and therefore $2NK$ runs of the model. However, according to the idea of Morris, it is possible to reduce this number to $(N+1)K$ [10].

Originally, the mean EE_j of all $(N+1)K$ parameters set was used to estimate the importance of parameter j. Since this approach might lead to compensation of positive and negative EEs, Campolongo et al. [11] proposed an alternative using:

$$\mu_j^* = \frac{1}{K}\sum_{l=1}^{K}\left|EE_j^{(l)}\right| \tag{5}$$

where K is the number of trajectories and $EE_j^{(l)}$ is the EE of parameter j within the l^{th} trajectory, and μ_j^* is the sensitivity measure of the j^{th} parameter.

In our work, we fixed $K = 20N$ trajectories which leads to a total number of $20N(N+1)$ parameter sets.

3. RESULTS AND DISCUSSION

Here we analyze the impact of all the parameters in the system (3). As a significant example, we report the results of the influence of parameters p_1 and p_2 (see Table 1 for details).

In the fixed interval $[0.0, 2.0]$, we search for 500 feasible values of p_1, randomly and uniformly distributed, as shown in Figure 1a. Then, we repeat the same procedure for p_2 in the interval $[0.0, 200.0]$ (Figure 1b). In the first case (Figure 1a), the parameter affect the solution slightly along both the x-axis and the y-axis, in particular, with a concentration of curves above the solution of Busse et al. [5] (represented by the thick line). Conversely, in the second case (Figure 1b) the shift along the y-axis is prevalent beneath the reference curve. Looking

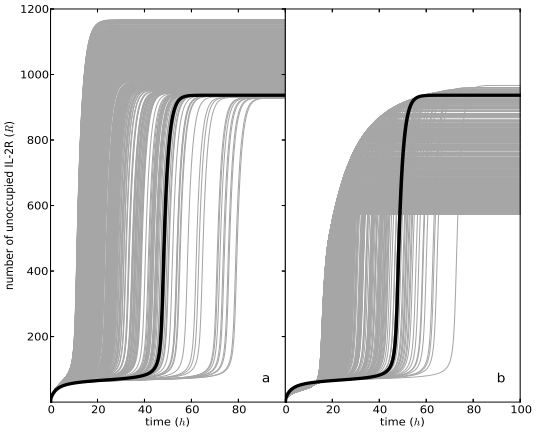


Figure 1. Influence on the solution of R of the system (3) (thick line) using 500 different values of a) p_1 and b) p_2.

more in detail, feasible parameters are found only in small subsets of the allowed intervals, specifically, in $[0.0, 0.7]$ for p_1 (Figure 2a) and in $[107.2, 200.0]$ for p_2 (Figure 2b). Moreover, it is interesting to observe that the values used for p_1 and p_2 in the work of Busse et al. [5] lie both in one of the extreme bounds of the feasible sets we obtained.

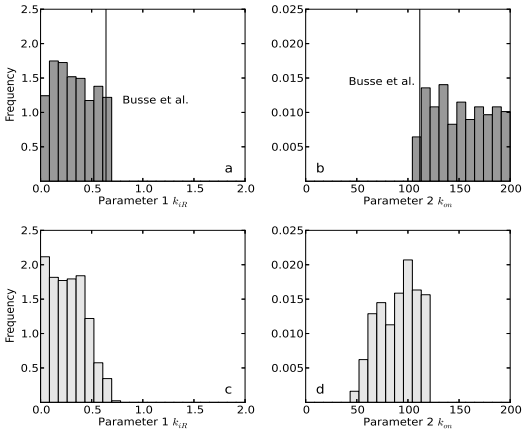


Figure 2. Frequency of the feasible parameters in the allowed intervals changing only a) p_1 and b) p_2. Comparison between modifying p_1 and p_2 together (light bars) and changing only c) p_1 and d) p_2.

In a second step, we now change p_1 and p_2 at the same time leaving all the other parameters fixed to the values of the work of Busse et al. [5]. In this case, the solution is affected along both axes but mainly along the y-axis reaching a plateau of nearly 3000 (Figure 3).

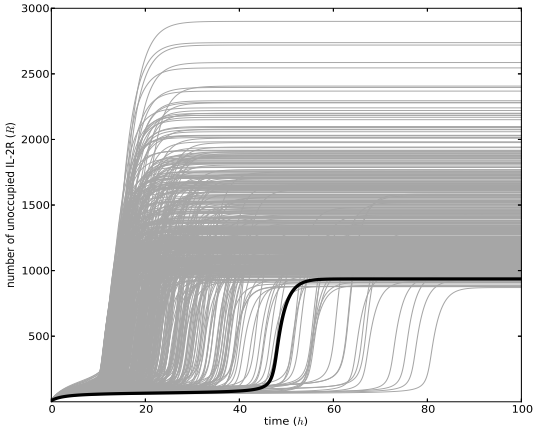


Figure 3. Influence on the solution of R of the system (3) (thick line) of 500 different values arising from changing parameters p_1 and p_2 together.

However, it is important that the behavior of the curve is not deducible from the previous two cases, when we changed only one parameter at once. The result is not representable as a combination of the two previous conditions, as clearly shown by the irregularity of the thin curves in Figure 3 and from the reached plateau that is almost three times higher than in the previous cases.

Moreover, replacing more parameters at the same time affects not only the solutions, but also the range in which is possible to find feasible parameters (Figure 2c-d). The feasible values for p_1 are found in a similar range compared to the previous case (Figure 2c). Conversely, for p_2 where we found a range completely different to the previous one, $[44.1, 120.0]$ (Figure 2d).

By applying the EE method to the system (3), it is possible to estimate the importance of every single parameter and rank them according to their influence on the model. As shown in Figure 4, parameters p_1 and p_8 perform major consequences on the system with respect to the others. On the contrary, p_7 and p_9 could be considered irrelevant in this model since their impact on the model output is much lower compared to the others. Therefore, we suggest to concentrate direct experimental measurements of parameters on p_1 and p_8 while excluding p_7 and p_9.

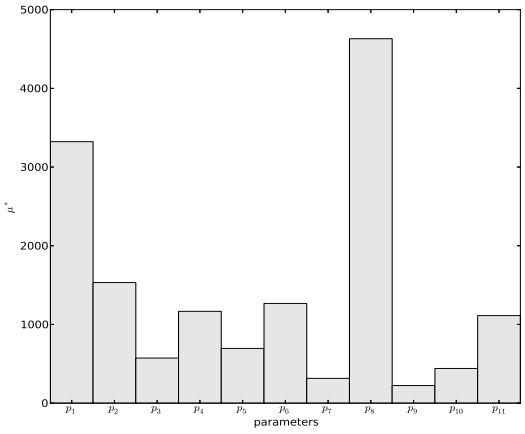


Figure 4. *Elementary Effects* μ^* of parameters of system (3).

Moreover, it is expected that p_1 and p_8 have an higher impact on the system since they appear only in the equation describing R (system (3)). These two parameters would not have the same influence if the EE method could be performed on the other variables (C, E) of the system.

Furthermore, aiming to evince which parameters are hidden in measurements any parameters inversion method can be focused on optimizing the most sensitive parameters, i. e. parameters with large EEs.

4. CONCLUSION

We showed how the outputs of a biological system are strictly related to the parameters present in the model. Moreover, we pointed out the importance of a prior sensitivity analysis, i.e. an *Elementary Effect* method, in order to identify the most relevant factors on which the model rely on and reduce the system complexity.

An improvement of this study could be the integration of IL-2 (C) and internalized IL-2R (E) in addition to the used unoccupied IL-2R (R). This will allow for an estimation of the overall impact of the parameters on the model and will therefore further enhance the efficiency of a subsequent parameter inversion methods.

5. ACKNOWLEDGMENTS

G. Dalmasso was supported by the Helmholtz Impulse and Networking Fund through the Helmholtz Interdisciplinary Graduate School for Environmental Research (HIGRADE).

6. REFERENCES

[1] B. D. Banerjee, "The influence of various factors on immune toxicity assessment of pesticide chemicals," *Toxicology Letters*, vol. 107, no. 1-3, pp. 21–31, Jun 30 1999, International Symposium on Health Aspects of Environmental and Occupational Pesticide Exposure, Dusseldorf, Germany, Sep 28-Oct 01, 1998.

[2] Z. Pancer and M. D. Cooper, "The evolution of adaptive immunity," *Annual Review of Immunology*, vol. 24, pp. 497–518, 2006.

[3] J. J. Hutton, A. G. Jegga, S. Kong, A. Gupta, C. Ebert, S. Williams, J. D. Katz, and B. J. Aronow, "Microarray and comparative genomics-based identification of genes and gene regulatory regions of the mouse immune system," *BMC Genomics*, vol. 5, Oct 25 2004.

[4] N. Bellomo and G. Forni, "Complex multicellular systems and immune competition: New paradigms looking for a mathematical theory," in *Multiscale Modeling of Developmental Systems*, Schnell, S and Maini, PK and Newman, SA and Newman, TJ, Ed., 2008, vol. 81 of *Current Topics in Developmental Biology*, pp. 485–502, 9th Biocomplexity Workshop, Bloomington, IN, May, 2006.

[5] D. Busse, M. de la Rosa, K. Hobiger, K. Thurley, M. Flossdorf, A. Scheffold, and T. Hoefer, "Competing feedback loops shape IL-2 signaling between helper and regulatory T lymphocytes in cellular microenvironments," *Proceedings of the National Academy of Sciences of the United States of America*, vol. 107, no. 7, pp. 3058–3063, Feb 16 2010.

[6] T. Aijo, S. Edelman, T. Lonnberg, A. Larjo, H. Kallionpaa, S. Tuomela, E. Engstrom, R. Lahesmaa, and H. Lahdesmaki, "An integrative computational systems biology approach identifies differentially regulated dynamic transcriptome signatures which drive the initiation of human t helper cell differentiation," *BMC Genomics*, vol. 13, no. 1, pp. 572, 2012.

[7] Z. Zi and E. Klipp, "SBML-PET: A systems biology markup language-based parameter estimation tool," *Bioinformatics*, vol. 22, no. 21, pp. 2704–2705, Nov 1 2006.

[8] P. Gonnet, S. Dimopoulos, L. Widmer, and J. Stelling, "A specialized ODE integrator for the efficient computation of parameter sensitivities," *BMC Systems Biology*, vol. 6, May 20 2012.

[9] R. P. Brent, "Some long-period random number generators using shifts and xors," *CoRR*, vol. abs/1004.3115, 2010.

[10] M. D. Morris, "Factorial sampling plans for preliminary computational experiments," *Technometrics*, vol. 33, no. 2, pp. 161–174, May 1991.

[11] F. Campolongo, J. Cariboni, and A. Saltelli, "An effective screening design for sensitivity analysis of large models," *Environmental Modelling & Software*, vol. 22, no. 10, pp. 1509–1518, Oct 2007.

PARAMETER ESTIMATION FOR STOCHASTIC BIOCHEMICAL PROCESSES: A COMPARISON OF MOMENT EQUATION AND FINITE STATE PROJECTION

Atefeh Kazeroonian[1], *Jan Hasenauer*[1,2], *and Fabian Theis*[1,2]

[1]Institute of Computational Biology, Helmholtz Center Munich,
Ingolstädter Landstr. 1, 85764 Neuherberg, Germany

[2]Department of Mathematics, University of Technology Munich,
Boltzmannstr. 3, 85747 Garching, Germany

{atefeh.kazeroonian,jan.hasenauer,fabian.theis}@helmholtz-muenchen.de

ABSTRACT

Many biochemical processes exhibit intrinsic stochastic fluctuations. These intrinsic fluctuations can be modeled using the chemical master equation (CME). The estimation of the parameters of the CME is challenging because the CME is a high or infinite dimensional system.

We compare two approaches currently used to estimate parameters of CMEs from population snapshot data. The first approach relies on a truncation of the CME, the finite state projection, and uses the data directly. The second method relies on moment equations – dynamical systems computing the moments of the CME solution – and merely uses the moments of the data. The second method is computationally more efficient, however, it cannot use all information contained in the data. In this manuscript, we assess the statistical power of the individual approaches and study moment equations of different order. Furthermore, we refine the likelihood function for the moment equation and introduce a novel validation method.

We performed a comparative study of the commonly used 3-stage model of gene expression. Using maximum likelihood estimates and a rigorous uncertainty quantification based on profile likelihoods, we show that the finite state projection approach is statistically more powerful than approaches based on moment equation. Nevertheless, even in case of partial observations, the first and second moments of the CME solution are highly informative and permit parameter identifiability. These findings, in combination with the novel tools for validation and uncertainty analysis, improve the insight into the problem class.

1. INTRODUCTION

In recent years, a multitude of studies have shown that many biochemical processes in prokaryotic and eukaryotic cells exhibit intrinsic stochastic fluctuations [1]. These fluctuations arise from low copy-number effects and are particularly significant for transcription and translation [2]. It is now known that these fluctuations are in many cases required for cellular function, e.g., for robust decision making on the population level [1].

The stochastic dynamics of biological processes can be described using continuous-time discrete-state Markov chains (CTMCs). The statistics of these Markov chains are governed by the chemical master equation (CME). Individual realizations of the process can be obtained via stochastic simulation algorithms (SSAs) [3, 4]. The stochastic process can be studied by analyzing statistics of many such realizations. Alternatively, the CME can be simulated using the finite state projection (FSP) method [5], which relies on truncation of the state space of the CME. While SSAs and the FSP are in principle capable of resolving all details of the dynamics of the CME, they impose a significant computational cost. This computational cost already becomes intractable for many small-scale systems. As an alternative, the method of moments (MM) [6, 7, 8] can be employed to capture the overall statistics of the process, such as mean and variance of individual species as well as covariances.

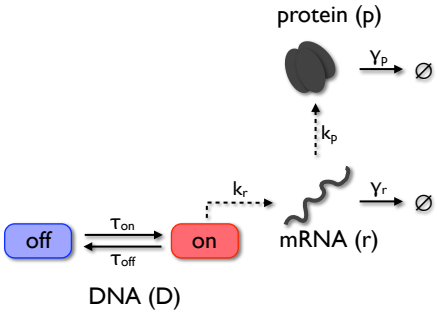
While the SSA, the FSP, and the MM all have advantages and disadvantages, a joint property is that they require accurate parameter values. The models and simulations are only predictive if good estimates of the reaction rates are available. Several estimation methods, relying on different models, were proposed (see, e.g., [9] and references therein), however, in most studies only the optimal parameter estimate has been considered, and the methods have not been compared. In this manuscript, we study the parameter estimates and confidence intervals obtained using FSP and MM. We present the individual likelihood functions and evaluate the informativeness using profile likelihoods. This is done for the widely used 3-stage model of gene expression [2], which is depicted in Figure 1.

2. METHODS

2.1. Modeling and simulation

2.1.1. Chemical master equation

The time evolution of the state $X = (X_1, \dots, X_{n_s})^{\mathrm{T}} \in \mathbb{N}_0^{n_s}$ of stochastic biochemical reaction networks is mostly described using CTMCs. The statistics of CTMCs are



Moment equation (order 1):

$$\dot{\mu}_{\mathrm{D_{off}}} = \tau_{\mathrm{off}}\mu_{\mathrm{D_{on}}} - \tau_{\mathrm{on}}\mu_{\mathrm{D_{off}}}$$
$$\dot{\mu}_{\mathrm{D_{on}}} = \tau_{\mathrm{on}}\mu_{\mathrm{D_{off}}} - \tau_{\mathrm{off}}\mu_{\mathrm{D_{on}}}$$
$$\dot{\mu}_{\mathrm{r}} = k_{\mathrm{r}}\mu_{\mathrm{D_{on}}} - \gamma_{\mathrm{r}}\mu_{\mathrm{r}}$$
$$\dot{\mu}_{\mathrm{p}} = k_{\mathrm{p}}\mu_{\mathrm{r}} - \gamma_{\mathrm{p}}\mu_{\mathrm{p}}$$

Figure 1. **Three-stage gene expression model.** (left) Schematic of the 3-stage gene expression model shows two DNA states (on, off), mRNAs and proteins. Transitions as well as synthesis and degradation reactions are shown as arrows. (right) Moment equations for means and variances of the individual species. The subscripts indicate the dependency, e.g., μ_r is the mean mRNA number.

governed by the CME. For a process with n_r chemical reactions,

$$R_k : \sum_{i=1}^{n_s} \nu_{ik}^- X_i \rightarrow \sum_{i=1}^{n_s} \nu_{ik}^+ X_i,$$

with reaction stoichiometries ν_k^-, ν_k^+, and $\nu_k = \nu_k^+ - \nu_k^-$, and reaction propensities $a_k(X, \theta)$, the CME is

$$\frac{\partial}{\partial t}p(x;t) = \sum_{\substack{k=1 \\ x \geq \nu_k^+}}^{n_r} a_k(x-\nu_k,\theta)p(x-\nu_k;t) - \sum_{k=1}^{n_r} a_k(x,\theta)p(x;t).$$

The solution of the CME depends on the parameters θ, which are for instance reaction rates.

The CME is defined for all reachable states $x \in \Omega \subset \mathbb{N}_0^{n_s}$, where n_s is the number of biochemical species. The set of reachable states Ω is in general very large, or infinite, rendering a direct solution of the full CME infeasible. Fortunately, the set of states with a significant probability mass is often small. This is exploited by the FSP, a direct method for approximating the solution of the CME [5] with pre-specified accuracy. Therefore, a subset Ω^{FSP} of the set of reachable states Ω is chosen. The time evolution of $p(x;t)$ with $x \in \Omega^{\mathrm{FSP}}$ is described by the CME, but influxes from states $x - \nu_k \notin \Omega^{\mathrm{FSP}}$ are removed. Probabilities $p(x;t)$ resulting from the simulation of this truncated system, which can be shown to be a lower bound for the actual probabilities of the CME, converge to the actual probabilities by growing Ω^{FSP} until the pre-specified accuracy is met.

A requirement for the application of the FSP is that the number of states with a significant probability mass is not too large. Novel algorithms can handle some million states [10]. Beyond this, the direct numerical simulation becomes infeasible.

2.1.2. Method of moments

In situations where the FSP is no longer applicable, the method of moments can be employed to approximate the solution of the CME [6]. The MM, also called moment equation, does not reproduce the exact solution of the CME. Instead, it computes the moments of $p(x;t)$, i.e. mean

$$\mu_i(t) = \sum_{x \in \Omega} x_i p(x;t),$$

variance

$$C_{ij}(t) = \sum_{x \in \Omega} (x_i - \mu_i(t))(x_j - \mu_j(t))p(x;t),$$

and higher-order moments [6]. The dynamics of the moments are governed by a set of ordinary differential equations (ODEs). Given that chemical reactions are at most bimolecular, the ODEs for the mean and the variance are

$$\frac{d\mu_i}{dt} = \sum_{k=1}^{n_r} \nu_{ik} \left(a_k(\mu,\theta) + \frac{1}{2} \sum_{l_1,l_2} \frac{\partial^2 a_k(\mu,\theta)}{\partial x_{l_1} \partial x_{l_2}} C_{l_1 l_2} \right),$$

$$\frac{dC_{ij}}{dt} = \sum_{k=1}^{n_r} \left(\nu_{ik} \sum_{l} \frac{\partial a_k(\mu,\theta)}{\partial x_l} C_{il} + \nu_{jk} \sum_{l} \frac{\partial a_k(\mu,\theta)}{\partial x_l} C_{jl} + \nu_{ik}\nu_{jk} \left(a_k(\mu,\theta) + \frac{1}{2} \sum_{l_1,l_2} \frac{\partial^2 a_k(\mu,\theta)}{\partial x_{l_1} \partial x_{l_2}} C_{l_1 l_2} \right) \right)$$
$$+ \sum_{k=1}^{n_r} \left(\nu_{ik} \sum_{l_1,l_2} \frac{\partial^2 a_k(\mu,\theta)}{\partial x_{l_1} \partial x_{l_2}} C_{i l_1 l_2} + \nu_{jk} \sum_{l_1,l_2} \frac{\partial^2 a_k(\mu,\theta)}{\partial x_{l_1} \partial x_{l_2}} C_{j l_1 l_2} \right),$$

in which $C_{il_1l_2}$ and $C_{jl_1l_2}$ are third order moments according to notation used in [6]. The governing equation for arbitrary moment orders can be found in [6, Equation (2.46)]. If all reactions are at most mono-molecular, the moment equation is closed, meaning that the evolution of moments of order m does not depend on moments of order greater than m. In this case, the moment equations are exact. If bimolecular chemical reactions are present, the moment equation ODEs are not closed, and the evaluation of a moment of order m requires the moments of order $m+1$ [6]. Moment closure techniques must be employed [11], and the resulting moments will only be an approximation of the true moments of the solution of the CME.

Moment equations are in general low-dimensional compared to the CME. Thus, they can generally be solved more efficiently. However, a drawback is that a finite number of moments does not allow the reconstruction of the underlying distribution $p(x;t)$. Hence, information is lost.

2.2. Parameter estimation

In this work, we considered population snapshot data $\mathcal{D}_k = \left\{ \left(\bar{Y}^{(s)}(t_k), t_k \right) \right\}_{s=1}^{S_k}$, $k = 1, \ldots, N$, obtained by sampling cells $s = 1, \ldots, S_k$ from the cell population and measuring one (or more) properties of these cells, e.g., using flow cytometry or microscopy. For notational simplicity, we assume that one observable, $\bar{Y} = h(X)$, can be measured. The observation function h describes the type of measurement; in the most simple case $h(X) = X_i$. The measurement is assumed to be noise-free as we later want to assess the informativeness of single-cell data vs. the moments.

Given a realization X at a certain time t_k, the probability of observing $\bar{Y}$ at time t_k is $p(y = \bar{Y}; x = X)$. The total probability to observe $\bar{Y}$ at time t_k is obtained by taking into account all possible realizations $X \in \Omega$ of the process. Given that the number of molecules is a discrete variable, this total probability is obtained by marginalizing over the state space Ω,

$$p(y; t_k, \theta) = \sum_{x \in \Omega} p(y; x)\, p(x; t_k, \theta),$$

where $p(x; t_k, \theta)$ is the solution of the CME. Bearing in mind that we do not consider any measurement noise, y is a deterministic function of x, $y = h(x)$, thus

$$p(y|x) = \begin{cases} 1 & \text{if } y = h(x) \\ 0 & \text{otherwise,} \end{cases}$$

so the sum simplifies to

$$p(y; t_k, \theta) = \sum_{\substack{x \in \Omega \\ h(x) = y}} p(x; t_k, \theta).$$

Following the argumentation above, the probability distribution $p(y; t_k, \theta)$ is the distribution from which the observations are drawn. Thus,

$$p(y = \bar{Y}^{(s)}(t_k)) = p(y; t_k, \theta), \quad s = 1, \ldots, S_k.$$

In the following, we compare two classes of likelihood functions for these data, namely an FSP-based likelihood function and a moment-based likelihood function with respect to their statistical power. As mentioned before, we do not consider any measurement noise in this comparison, but the inclusion of noise in the presented procedure would be rather straightforward.

2.2.1. FSP-based estimation

As outlined earlier, for CTMCs with a small effective state space, the FSP can be used to approximate the solution of the CME for a given parameter set θ. Using this approximation of the probability distribution of the hidden state, $p(x; t, \theta)$, and the corresponding approximation of the probability distribution of the observable, $p(y; t, \theta)$, the likelihood of the stochastic process,

$$\mathcal{L}_{\mathcal{D}}^{\text{FSP}}(\theta) = c \prod_{k=1}^{N} \prod_{s=1}^{S_k} p(y = \bar{Y}^{(s)}(t_k); t_k, \theta),$$

can be evaluated. Basically, the probabilities are evaluated and multiplied for all observed states. The constant c depends only on the data and can be neglected for optimization purposes. For a detailed introduction of this FSP-based likelihood function, we refer to [12, 13]. Given the FSP-based likelihood function, the estimation problem can be formulated. The FSP-based maximum likelihood (ML) estimation problem is:

$$\begin{aligned} &\underset{\theta}{\text{maximize}} \ \log \mathcal{L}_{\mathcal{D}}^{\text{FSP}}(\theta) \\ &\text{subject to } \Sigma^{\text{FSP}}(\theta), \end{aligned}$$

in which $\Sigma^{\text{FSP}}(\theta)$ denotes the finite-dimensional ODE model resulting from the FSP of the CME on the subset Ω^{FSP}. To reduce numerical problems, the problem is formulated using the log-likelihood function $\log \mathcal{L}^{\text{FSP}}_{\mathcal{D}}(\theta)$. Furthermore, we optimize the logarithm of the parameters $\xi = \log_{10}(\theta)$ to ensure positivity and improve the performance of the optimization routines. The optimal solution of the FSP-based ML estimation problem is the parameter vector for which the likelihood of observing the single cell data is maximized. This estimator uses all available information.

2.2.2. Moment-based estimation

For many processes the approximation of the CME solution using the FSP is not feasible because the number of states with non-negligible probability is too large. In such cases, the moment equation can be employed to approximate the statistics of the CME solution. To employ moment equations for parameter estimation, the statistics of the snapshots are computed, e.g., mean and variance,

$$\bar{\mu}_y(t_k) = \frac{1}{S_k}\sum_{s=1}^{S_k} \bar{Y}^{(s)}(t_k),$$

$$\bar{C}_{yy}(t_k) = \frac{1}{S_k}\sum_{s=1}^{S_k} \left(\bar{Y}^{(s)}(t_k) - \bar{\mu}_y(t_k)\right)^2.$$

These measured moments are compared to moments predicted by the model and the observation function $h(x)$. Since the sample sizes S_k are often quite large – for flow cytometry often in the order of 10^4 – it follows from the central limit theorem that the empirical moments, e.g., $\bar{\mu}_y(t_k)$ and $\bar{C}_{yy}(t_k)$, are almost normally distributed around the true moments [14]. Hence, a normal error model is assumed,

$$\mathcal{L}^{\text{MM}}_{\mathcal{D},\mu_y}(\theta) = \prod_{k=1}^{N} \mathcal{N}\left(\mu_y(t_k,\theta)|\bar{\mu}_y(t_k), \sigma^2_{\bar{\mu}_y}(t_k)\right),$$

$$\mathcal{L}^{\text{MM}}_{\mathcal{D},C_{yy}}(\theta) = \prod_{k=1}^{N} \mathcal{N}\left(C_{yy}(t_k,\theta)|\bar{C}_{yy}(t_k), \sigma^2_{\bar{C}_{yy}}(t_k)\right),$$

where $\mathcal{N}(\cdot|\mu,\sigma^2)$ is the probability density of the normal distribution. Such a likelihood function can be derived for every moment predicted by the model, e.g., also the third and fourth order central moments. Clearly, the consideration of additional, non-redundant moments provides additional information about the model parameters as the individual likelihood functions are multiplied, e.g., if mean and variance are employed then a reasonable likelihood function is

$$\mathcal{L}^{\text{MM}}_{\mathcal{D}}(\theta) = \mathcal{L}^{\text{MM}}_{\mathcal{D},\mu_y}(\theta) \cdot \mathcal{L}^{\text{MM}}_{\mathcal{D},C_{yy}}(\theta).$$

Unfortunately, also the computational complexity of simulating the moment equations increases with each additional moment considered in the model.

The likelihoods $\mathcal{L}^{\text{MM}}_{\mathcal{D},\mu_y}(\theta)$, $\mathcal{L}^{\text{MM}}_{\mathcal{D},C_{yy}}(\theta)$ and those for the higher-order moments require information about the error variance of the respective empirical estimator, e.g., $\sigma^2_{\bar{\mu}_y}$ for $\bar{\mu}_y(t_k)$ and $\sigma^2_{\bar{C}_{yy}}$ for $\bar{C}_{yy}(t_k)$. The variance of the estimators for the first and second order moments can be found in [14]. For third and higher-order moments the calculation of these estimators become increasingly complex, and we did not find respective results in the literature. To circumvent the analytical derivation, we propose to estimate the variance of the empirical estimators using non-parametric bootstrapping [15]. This approach employs a two-step procedure. At first, a sample of size S_k is drawn from $\{\bar{Y}^{(s)}(t_k)\}_{s=1}^{S_k}$ (all $\bar{Y}^{(s)}(t_k)$ have probability $\frac{1}{S_k}$) and the moments of this artificial sample are evaluated. This step is repeated a large number of times, in general more than one thousand times, yielding a large sample for each moment of interest. Therefore, the variance of each moment can easily be computed from the corresponding sample. This sample variance is a reliable measure for the uncertainty, if $S_k \gg 1$. It does not require any distribution assumption for $p(y;t_k,\theta)$ and is easily applicable to any higher-order moments.

Given the likelihood function $\mathcal{L}^{\text{MM}}_{\mathcal{D}}(\theta)$, which is the product of the likelihood functions for the moments of interest, the moment-based ML estimation problem,

$$\begin{aligned} &\underset{\theta\in\mathbb{R}_+^{n_\theta}}{\text{maximize}} \ \log \mathcal{L}^{\text{MM}}_{\mathcal{D}}(\theta) \\ &\text{subject to } \Sigma^{\text{MM}}(\theta), \end{aligned}$$

can be formulated. $\Sigma^{\text{MM}}(\theta)$ is the model used to simulate the moment equations for the moments of interest.

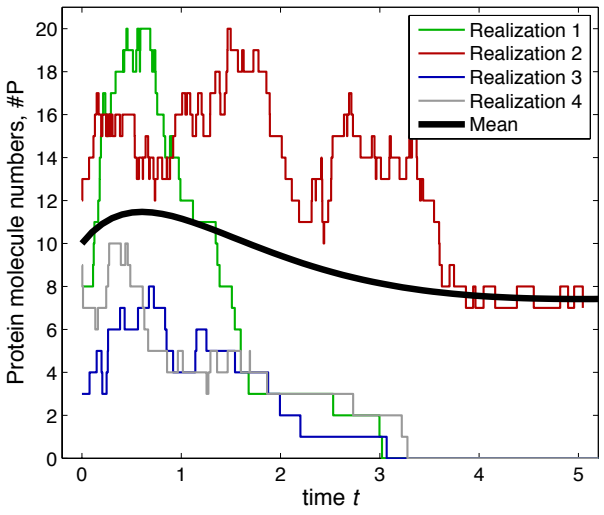
2.2.3. Identifiability and uncertainty analysis

As the measurement data are limited and potentially noise corrupted, the parameters can in general not be estimated precisely. To assess the remaining parameter uncertainty and the practical identifiability, we use profile likelihoods [16]. Given the likelihood function $\mathcal{L}_{\mathcal{D}}(\theta)$, the profile likelihood of parameter θ_i is

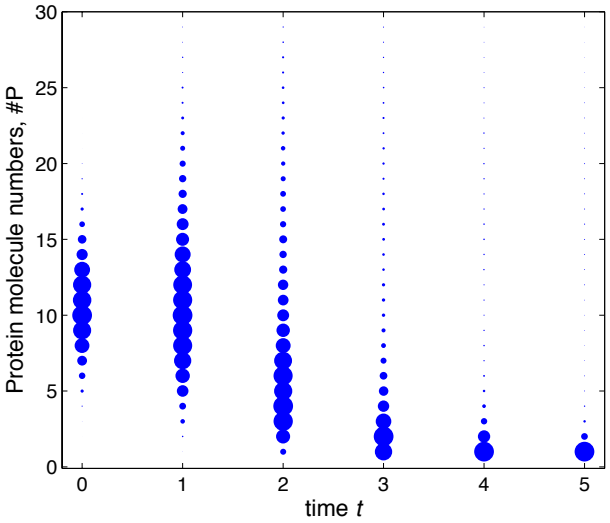
$$\text{PL}(\theta_i) = \max_{\theta_{j\neq i}} \mathcal{L}_{\mathcal{D}}(\theta).$$

This profile likelihood $\text{PL}(\theta_i)$ is the maximal likelihood for a given value of θ_i. Using the profile likelihood, the likelihood ratio $R(\theta_i) = \text{PL}(\theta_i)/\mathcal{L}_{\mathcal{D}}(\hat{\theta})$ can be evaluated, in which $\hat{\theta}$ is the ML estimate. The likelihood ratio R is one at the globally optimal point $\hat{\theta}_i$ and approaches zero for large $|\theta_i - \hat{\theta}_i|$ if the parameter is identifiable. The area under $\text{PL}(\theta_i)$ provides a reasonable measure for the uncertainty of parameter θ_i. For further details, we refer to [16, 17].

In the following, we employ profile likelihoods to assess the information content of the moments of the data in comparison with that of the full distribution of data. More information will result in many identifiable parameters and small parameter uncertainties.



(a) **Four stochastic realizations of the 3-stage model of gene expression.**



(b) **Population snapshot data used for parameter estimation.**

Figure 2. **Dynamics of the 3-stage model of gene expression.** (a) Time-dependent protein number in four representative cells together with the population mean. (b) Population snapshot data obtained by sampling single cell trajectories. The size of the markers in (b) is proportional to the number of observed cells with the corresponding protein number. Due to the long tail of the distribution, the mode of the data seen in (b) differs significantly from the mean of the data depicted in (a).

3. RESULTS AND DISCUSSION

3.1. Parameter estimation for the 3-stage model of gene expression

In this section, we compare the performance of previously mentioned estimation methods, namely, FSP-based and MM-based parameter estimates, using the common 3-stage model of gene expression [2]. A schematic of the process and the corresponding moment equations for mean and variance are shown in Figure 1. The model has six parameters: the transition rate of DNA into the on-state (τ_{on}), the transition rate of DNA into the off-state (τ_{off}), the transcription rate in the on-state (k_{r}), the rate of mRNA degradation (γ_{r}), the translation rate (k_{p}), and the rate of protein degradation (γ_{p}). In the following, we study the problem of estimating these rates from protein measurements. Therefore, we generate artificial data

$$\mathcal{D}_k = \left\{ \left(\bar{Y}^{(s)}(t_k), t_k \right) \right\}_{s=1}^{10^5}, \quad k = 1, \ldots, 10,$$

with $t_k = k$ and $\bar{Y}$ being the number of proteins. For the generation of the artificial data, the parameter vector

$$\begin{aligned} \theta^{\text{true}} &= (\tau_{\text{on}}, \tau_{\text{off}}, k_{\text{r}}, \gamma_{\text{r}}, k_{\text{p}}, \gamma_{\text{p}})^{\text{T}} \\ &= (0.05, 0.05, 5, 1, 4, 1)^{\text{T}} \end{aligned}$$

is used. We refer to this parameter vector θ^{true} as the true parameter vector in the following. Also, no measurement noise is considered in the generation of the data. In the initial state, mRNA and protein numbers follow a Poisson distribution with mean 4 and 10, respectively. The probability to be in the DNA on-state is 0.7. Figure 2 depicts sample paths of the model (Figure 2(a)) as well as the snapshot data (Figure 2(b)) used for parameter estimation. Using these data we estimate $\theta = (\tau_{\text{on}}, \tau_{\text{off}}, k_{\text{r}}, \gamma_{\text{r}}, k_{\text{p}}, \gamma_{\text{p}})^{\text{T}}$.

For FSP-based and moment-based likelihood functions the maximum likelihood estimates are computed and the parameter uncertainty is evaluated. For the moment-based likelihood function we employed different moment orders. The uncertainty of the moments has been determined using the non-parametric bootstrapping approach introduced before.

Figure 3 depicts the model simulation for the ML estimates for the different likelihood functions along with the data. It is clear that for all ML estimates we observe a good agreement with the data used for the estimation. To validate the ML estimates, we employed the higher-order moments of the data, which have not been used for the parameter estimation. We find that all ML estimates, which were obtained using at least the mean and the variance, successfully predict the higher-order moments not used to obtain the ML estimates. Only the ML estimate computed merely from the mean of the data fails. Thus, the information contained in the mean is insufficient. This is confirmed by the profile likelihoods shown in Figure 4, which show that all likelihood functions establish identifiability, except the moment-based likelihood function of order 1. A careful comparison of the profile likelihoods shows that the uncertainty in the estimation of the parameters decreases as more information (more moments) are used. Since the FSP-based likelihood function makes use of all the information, the resulting parameter uncertainties are minimal. If the moment order is increased, the confidence intervals for moment-based likelihood function also become more narrow, however even for moment order 4, the result of the FSP remains superior. Note that for all likelihood functions, the true parameters are con-

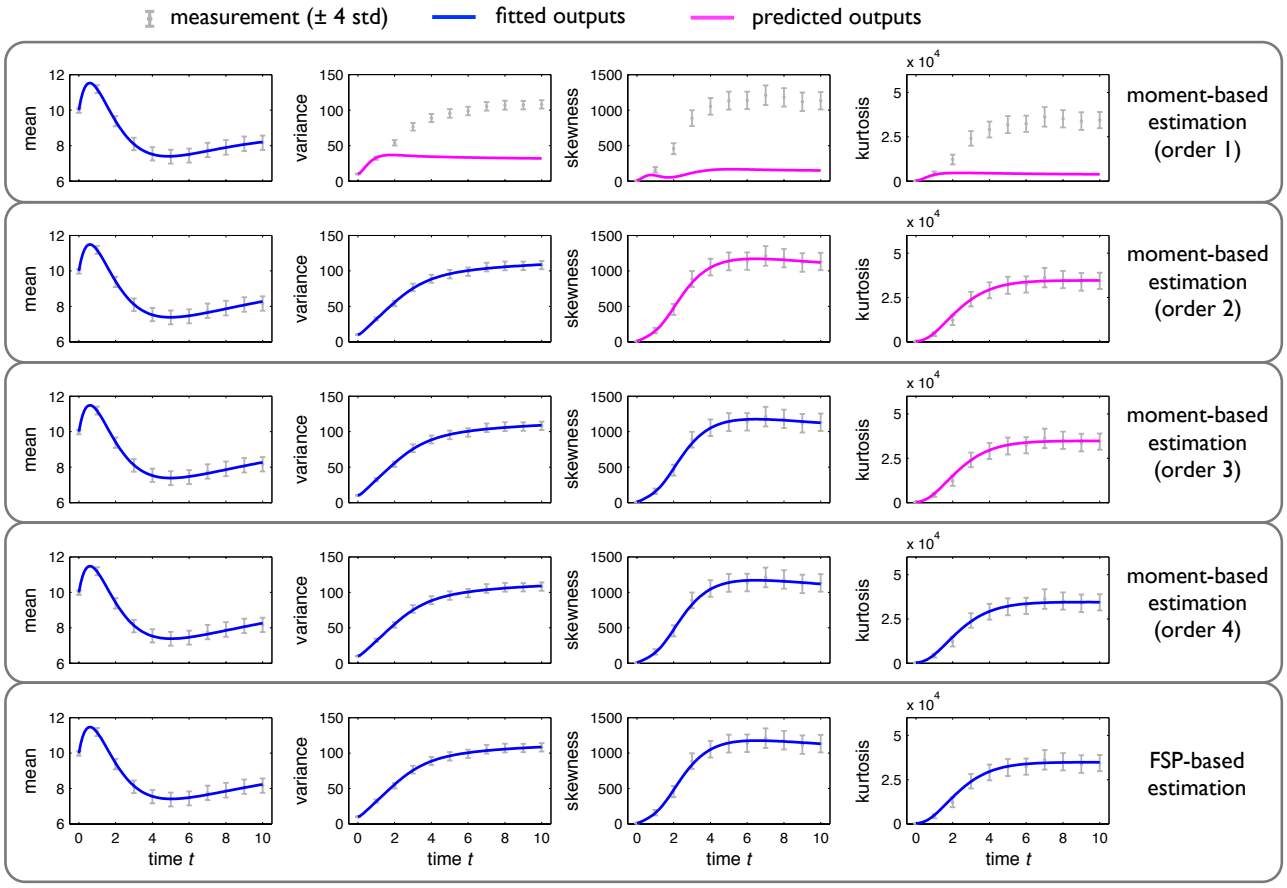


Figure 3. **Model-data comparison for ML estimates obtained using different likelihood functions.** ML estimation has been performed using moment-based likelihood functions of different orders (order 1: mean; order 2: mean and variance; order 3: mean, variance and skewness; and order 4: mean, variance, skewness and kurtosis) and the FSP-based likelihood function. Gray error bars show the mean and 4-σ intervals ($[\mu-4\sigma, \mu+4\sigma]$) of the measurement data. For the different ML estimates the fit is illustrated by showing the model output (blue lines, —) and the measurement data (grey error bars). All models describe the respective data well. To assess the predictive power of the model, the ML estimates are used to predict the higher-order moments (magenta lines, —) which have not been employed for the parameter estimation. The ML estimate computed using moment-based estimation of order 1 fails to provide good prediction, while already information about mean and variance (order 2) is sufficient to obtain a predictive model.

tained in the 95% confidence intervals constructed from the profile likelihoods (not shown).

3.2. Discussion

The computational complexity of the simulation of CTMCs renders the estimation of their parameters challenging. Different methods have been proposed to circumvent this complexity, among other the moment equations [18, 9, 14]. In this work, we evaluate the information contained in the moments of measurement data with respect to parameter estimation (by employing moment-based likelihood function) and compare it with the complete information contained in population snapshot data (by employing FSP-based likelihood function). The practical identifiability and the uncertainty of the parameter estimates are assessed using profile likelihoods. To the best of our knowledge, this is the first profile likelihood-based uncertainty analysis for stochastic processes, probably because the evaluation of the likelihood function is computationally often infeasible. This is not the case if a moment-based estimation is employed.

As a case study, we consider the widely used 3-stage model of gene expression [2]. For this model, we show that measurements of the mean expression do not in general ensure identifiability, but rather that measurements of the variance are required. This is consistent with results by Munsky *et al.* [18] for the two-stage model of gene expression. Information about third and fourth order moments can decrease the uncertainty further, however this reduction is often insignificant. The full information contained in the data, which is exploited by the FSP-based estimation, remains out of reach for the MM-based estimation approach.

Although the FSP-based likelihood function is statistically more powerful, parameter estimation based on the moment equation is the method of choice for processes,

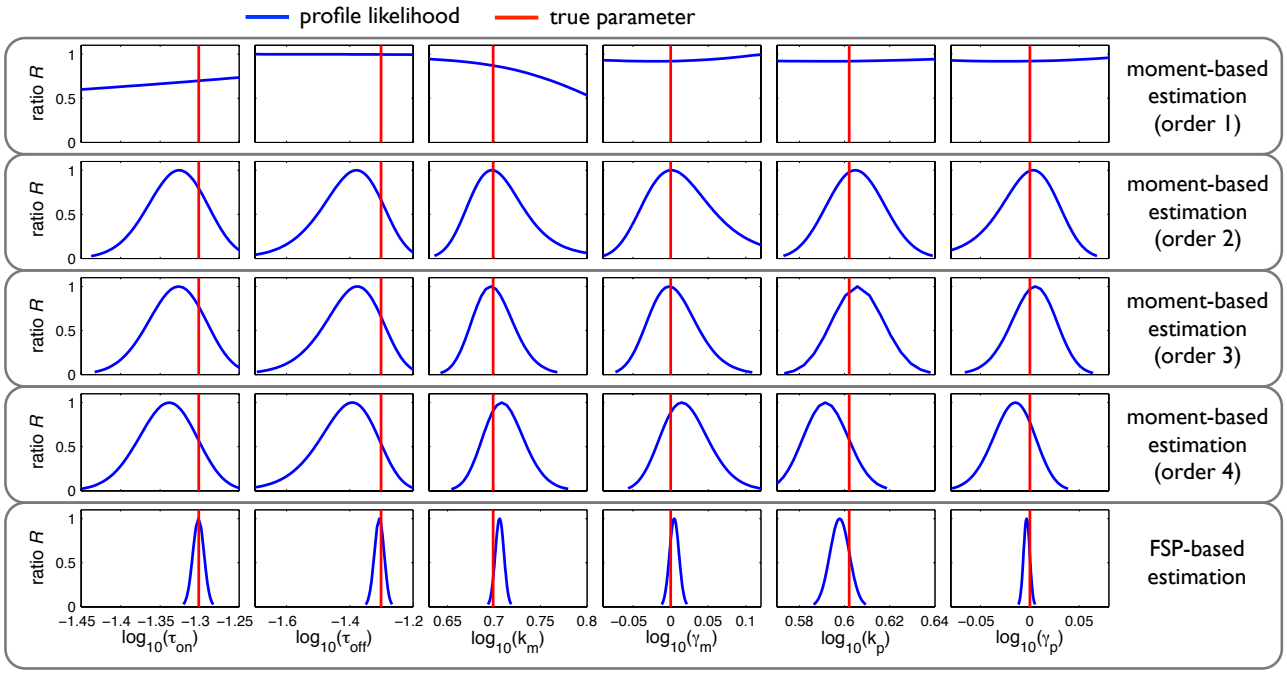


Figure 4. **Parameter uncertainty for different likelihood functions.** The parameter uncertainty and parameter identifiability has been evaluated for moment-based likelihood functions of different orders (order 1: mean; order 2: mean and variance; order 3: mean, variance and skewness; and order 4: mean, variance, skewness and kurtosis) and the FSP-based likelihood function. The profile likelihoods (blue lines, —) indicate that the measurements of the mean do not carry enough information to identify the parameters. Information about mean and variance ensures identifiability, and the uncertainty is slightly reduced if additional moments are used. The FSP-based likelihood function, which exploits all information contained in the data, yields the smallest uncertainties. All confidence intervals (not shown), derived from likelihood profiles, contain the true parameter values (red lines, —), which indicates consistency.

in particular, if the FSP is infeasible. Furthermore, parameter estimation using the moment equation is more efficient. The parameter estimation using the moment equation of order 2 is roughly 30 times faster than the parameter estimation using the FSP. However, it remains to be studied how moment closures, which are required for systems including bimolecular reactions, influence the parameter estimation. If a bias is introduced, as we expect, it should be analyzed how a refinement of the moment equation, e.g., the conditional moment equation [19], can be used to improve the results.

Beyond the profile likelihood-based evaluation of the information encoded in the moments, we introduced a non-parametric bootstrapping approach to evaluate the uncertainty of the empirical estimates of the moments. This approach allows for the construction of likelihood function without additional distribution assumptions. Furthermore, we illustrated how the higher-order moments, which have not been used for parameter estimation, can be used for model validation. This approach is attractive, as models can basically be fitted and validated on the same dataset.

4. AUTHOR'S CONTRIBUTIONS

AK and JH developed the method and analyzed the 3-stage model of gene expression. JH and FJT devised the project. AK, JH and FJT wrote, read and approved the final manuscript.

5. ACKNOWLEDGEMENTS

The authors acknowledge financial support by the European Union within the ERC grant 'LatentCauses' and the BMBF grant 'Virtual Liver' (grant-nr. 315752). The authors would also like to thank Justin Feigelman and Sabine Hug for proofreading the manuscript.

6. REFERENCES

[1] A. Eldar and M. B. Elowitz, "Functional roles for noise in genetic circuits," *Nat.*, vol. 467, no. 9, pp. 1–7, Sept. 2010.

[2] V. Shahrezaei and P. S. Swain, "Analytical distributions for stochastic gene expression," *Proc. Natl. Acad. Sci. U S A*, vol. 105, no. 45, pp. 17256–17261, Nov. 2008.

[3] D. T. Gillespie, "Exact stochastic simulation of coupled chemical reactions," *J. Phys. Chem.*, vol. 81, no. 25, pp. 2340–2361, Dec. 1977.

[4] H. E. Samad, M. Khammash, L. Petzold, and D. Gillespie, "Stochastic modelling of gene regulatory networks," *Int. J. Robust Nonlinear Control*, vol. 15, no. 15, pp. 691–711, Oct. 2005.

[5] B. Munsky and M. Khammash, "The finite state projection algorithm for the solution of the chemical master equation," *J. Chem. Phys.*, vol. 124, no. 4, pp. 044104, Jan. 2006.

[6] S. Engblom, "Computing the moments of high dimensional solutions of the master equation," *Appl. Math. Comp.*, vol. 180, pp. 498–515, 2006.

[7] J. P. Hespanha, "Modeling and analysis of stochastic hybrid systems," *IEE Proc. Control Theory & Applications,* Special Issue on Hybrid Systems, vol. 153, no. 5, pp. 520–535, 2007.

[8] J. Ruess, A. Milias, S. Summers, and J. Lygeros, "Moment estimation for chemically reacting systems by extended Kalman filtering," *J. Chem. Phys.*, vol. 135, no. 165102, Oct. 2011.

[9] P. Milner, C. S. Gillespie, and D. J. Wilkinson, "Moment closure based parameter inference of stochastic kinetic models," *Stat. Comp.*, 2012.

[10] M. Mateescu, V. Wolf, F. Didier, and T. Henzinger, "Fast adaptive uniformisation of the chemical master equation," *IET. Syst. Biol.*, vol. 4, no. 6, pp. 441–452, 2010.

[11] A. Singh and J. P. Hespanha, "Approximate moment dynamics for chemically reacting systems," *IEEE Trans. Autom. Control*, vol. 56, no. 2, pp. 414–418, Feb. 2011.

[12] J. Hasenauer, N. Radde, M. Doszczak, P. Scheurich, and F. Allgöwer, "Parameter estimation for the CME from noisy binned snapshot data: Formulation as maximum likelihood problem," Extended abstract at *Conf. of Stoch. Syst. Biol.*, Monte Verita, Switzerland, July 2011.

[13] T. Nüesch, "Finite state projection-based parameter estimation algorithms for stochastic chemical kinetics," Master thesis, Swiss Federal Institute of Technology, Zürich, 2010.

[14] C. Zechner, J. Ruess, P. Krenn, S. Pelet, M. Peter, J. Lygeros, and H. Koeppl, "Moment-based inference predicts bimodality in transient gene expression," *Proc. Nati. Acad. Sci. U S A*, vol. 109, no. 21, pp. 8340–8345, May 2012.

[15] T. J. DiCiccio and B. Efron, "Bootstrap confidence intervals," *Statist. Sci.*, vol. 11, no. 3, pp. 189–228, 1996.

[16] A. Raue, C. Kreutz, T. Maiwald, J. Bachmann, M. Schilling, U. Klingmüller, and J. Timmer, "Structural and practical identifiability analysis of partially observed dynamical models by exploiting the profile likelihood," *Bioinf.*, vol. 25, no. 25, pp. 1923–1929, May 2009.

[17] W. Q. Meeker and L. A. Escobar, "Teaching about approximate confidence regions based on maximum likelihood estimation," *Am. Stat.*, vol. 49, no. 1, pp. 48–53, Feb 1995.

[18] B. Munsky, B. Trinh, and M. Khammash, "Listening to the noise: random fluctuations reveal gene network parameters," *Mol. Syst. Biol.*, vol. 5, no. 318, Oct. 2009.

[19] J. Hasenauer, V. Wolf, A. Kazeroonian, and F. J. Theis, "Method of conditional moments (MCM) for the chemical master equation," *submitted to the Journal of Mathematical Biology*, 2012.

CLASSIFICATION OF SPECIES TO HIGHER TAXA BASED ON ANALYSIS OF DNA BARCODES – A BIRD EXAMPLE

Denisa Maderankova[1] and Ivo Provaznik[2]

[1]Department of Biomedical Engineering, Brno University of Technology
Technicka 12, 616 00 Brno, Czech Republic
[2]International Clinical Research Center - Center of Biomedical Engineering, St. Anne's University Hospital Brno, Brno, Czech Republic
maderankova@feec.vutbr.cz, provaznik@feec.vutbr.cz

ABSTRACT

DNA barcoding based on mitochondrial gene analysis is a modern method for species identification. We tested three variations of a novel method for species classification into families and orders based on the nucleotide density of the mtDNA barcode. We verified the methods on datasets of bird's barcode sequences. The reference database of species families was created from South American bird species. North American, European, and Asian birds of the same families but mostly different species were classified effectively nearly 88.13 percent into families and 99.51 percent into orders in best case for the third variation of the method.

1. INTRODUCTION

There is a great need for species identification and taxonomic classification tools. The number of species currently living on Earth is based on the opinion of taxonomic experts and is estimated as 3–100 million prokaryotes and eukaryotes (4 percent of them described) [1]. Traditional morphological taxonomy methods are unable to process such a huge number and diversity of species, particularly prokaryotes. Species identification and classification is a task for which molecular taxonomy based on the investigation of DNA sequences is more convenient. The molecular taxonomy of eukaryotes studies nuclear DNA and mitochondrial or plastid DNA.

The mitochondrial genome is a popular marker of molecular diversity in a variety of scientific fields, including population genetics, phylogenetics, phylogeography, and molecular ecology [2, 3]. An easy extraction from the cell, no introns, no insertions and deletions, and short intergenetic regions are great advantages of mtDNA. Its usage in determining molecular diversity is based on three assumptions, such as clonality [4], neutrality of mutations or slightly deleterious mutations [5], and constant evolutionary rate [6] that are currently in dispute [7].

Despite the disputation, mitochondrial DNA is still frequently used for many types of studies. In recent years, Paul Hebert proposed DNA barcoding as a method for species identification through mtDNA [8]. Identification based on mtDNA sequences of contemporary species suffers less from the limited validity of clonality, neutrality, and constant evolutionary rate, but it still has its disadvantages [9, 10]. DNA barcoding identifies species through genome analysis, allowing the analysis of even unknown tissues samples, microscopic species, closely related species, and morphologically cryptic species. DNA barcoding usually uses 648 base-pair region of cytochrome c oxidase subunit 1 gene (*cox1*). A Barcode of Life Data Systems (BOLD) database has been launched and provides free access to data and some analysis tools. BOLD database currently comprises more than 171,000 of species barcode records and almost 2 million of specimen's records.

There are many interesting papers focused on species classification based on DNA barcoding and other genetic markers like [11] which compares phylogenetic and statistical classification methods or [12] evaluating many classification programs. Phylogenetic tree construction and homology sequence alignment is used for identification of query sequence in BOLD database.

We were interested if the DNA barcode is characteristic not only for species but also for higher taxa like as families and orders. We used nucleotide density vectors (ND) of DNA barcodes as references for classification into families based on comparison of a sample sequence with reference sequences.

2. METHODS

2.1. Data

Our survey comprises these DNA barcoding projects: Birds of Argentina – Phase I (BARG), DNA Barcoding Korean Birds (KBBI), Birds of North America (TZBNA), Birds of North America – Canadian geese (BNACA), Birds of North America – Canadian passerines (BNABS), Birds of North America – General sequences (BNAUS), Birds of the eastern Palearctic (BEPAL), Birds of Scandinavia – Swedish birds (SWEBI), and Birds of Scandinavia – Norwegian birds (NORBI).

All sequences were downloaded as FASTA files from BOLD database. Only sequences longer than 600 base pairs (bp) were included in the datasets. Significantly shorter sequences than usually 648 bp long barcodes

could negatively influence the calculation of the references. The BARG project was selected for creating reference databases of nucleotide densities. The other projects were used only for classification analysis.

Sequences of all projects were at first sorted according recent scientific classification. The scientific classification based on morphological and behavioral resemblance of species has limitations, and the classification of some species is still disputed. This classification may differ from classification based on the recent molecular phylogeny. However, the use of the scientific classification to create references misrepresents results only in a few cases.

2.2. Nucleotide density

Nucleotide density (ND) is a simple and efficient numerical representation of a symbolic DNA sequence. It expresses an average occurrence of nucleotides in a defined region of the sequence. When calculating the nucleotide densities, binary indicator vectors $u_A[n]$, $u_C[n]$, $u_G[n]$, and $u_T[n]$ are created from the symbolic sequence as first step. The binary indicator vectors contain the value 1 when the corresponding nucleotide exists at position n in the sequence or the value 0 when it does not exist. To eliminate the effect of the beginning and the ending of the sequence, $W/2$ number of zeros is added to the sides of the binary indicator vectors before ND calculation. W is a size of a moving window where the calculation is performed. Then, the nucleotide densities are calculated for each type of nucleotide according to:

$$d_x[n] = \frac{\sum_{i=n-W/2}^{n+W/2} u_X[i]}{W}, n = 1,...,N, \qquad (1)$$

where N is the length of the sequence and X is one of the four nucleotides. The size of the moving window W has to be odd number because the position n is in the center of the moving window. The size of the moving window has to be set according to the desired difference resolution and sequence length.

2.3. Reference databases of bird's families

The reference databases were created from the DNA barcodes of the BARG project. There are 1588 sequences in the file. A dataset for the reference database creation contained only bird's family sequences in which at least three different species were included in the BARG project. All specimens of a species were included into the dataset. The completed dataset contained 1450 specimens of 442 species which belong to 38 families of 18 orders.

Reference databases of NDs were created with three different variations of method. As the nucleotide densities are calculated in moving window, its size can affect the results. There were created reference databases for different sizes of moving window from 5 to 29 nucleotides (odd sizes only) with each method variation.

The first variation (VRC) used nucleotide densities of consensus sequence as reference which was obtained from globally aligned barcode sequences of a bird's family. The second variation (VRM) used median values of separately calculated and correlated NDs of each sequence belonging to a particular bird's family. The third variation (VRA) used average values of separately calculated and correlated NDs.

2.4. Method of classification

All sequences from a dataset for classification analysis were compared with the references from databases of each method variation and moving window size. For comparison with the VRC references, the analyzed sequences were independently aligned with reference consensus sequences, then the NDs were calculated, and finally the Euclidean distances between the NDs of the analyzed sequences and the consensus references were calculated. The results of comparison with the VRM and the VRA references differ. Analyzed sequences were not aligned with consensus, but their nucleotide densities were calculated. Then, correlation function (signal processing method) was used to find best mutual position of the densities. Finally, the Euclidean distances between the NDs were calculated.

For all methods, the analyzed sequences were classified into families according to minimal Euclidean distance values.

3. RESULTS

3.1. Verification of method

All reference databases were verified by classification of sequences from the BARG project. These sequences were used for creating the databases so high effectiveness of their classification is expected if the used method variations for creating the references are suitable. The sequences were classified into the reference families. Efficiencies of the classification to the families were calculated as the ratio of number of sequences to number of correctly classified sequences. If there is more than one family of particular order among the references, the efficiency for the order classification can be also calculated. If there is only one family the efficiency values for order and family are equal.

Results of verification differ for each of the method variation. The worst results belong to the VRC and the best results were obtained from the VRA. The families with the poorest VRC classification results were significantly better classified with the VRM and the VRA references. The left side of the Figure 1 shows graphical representation of the classification efficiencies separately for each family. The Figure 2a shows weighted averages of classification efficiencies for all families and orders dependent on moving window size.

The major of the BARG sequences belongs to the *Passeriformes* order. There are 17 families of 945 sequences in total. The *Passeriformes* order is very broad with many closely related families. There are also species with disputed family classification. The order classification is almost perfect for all methods. The family classification was for the VRC method often incorrect.

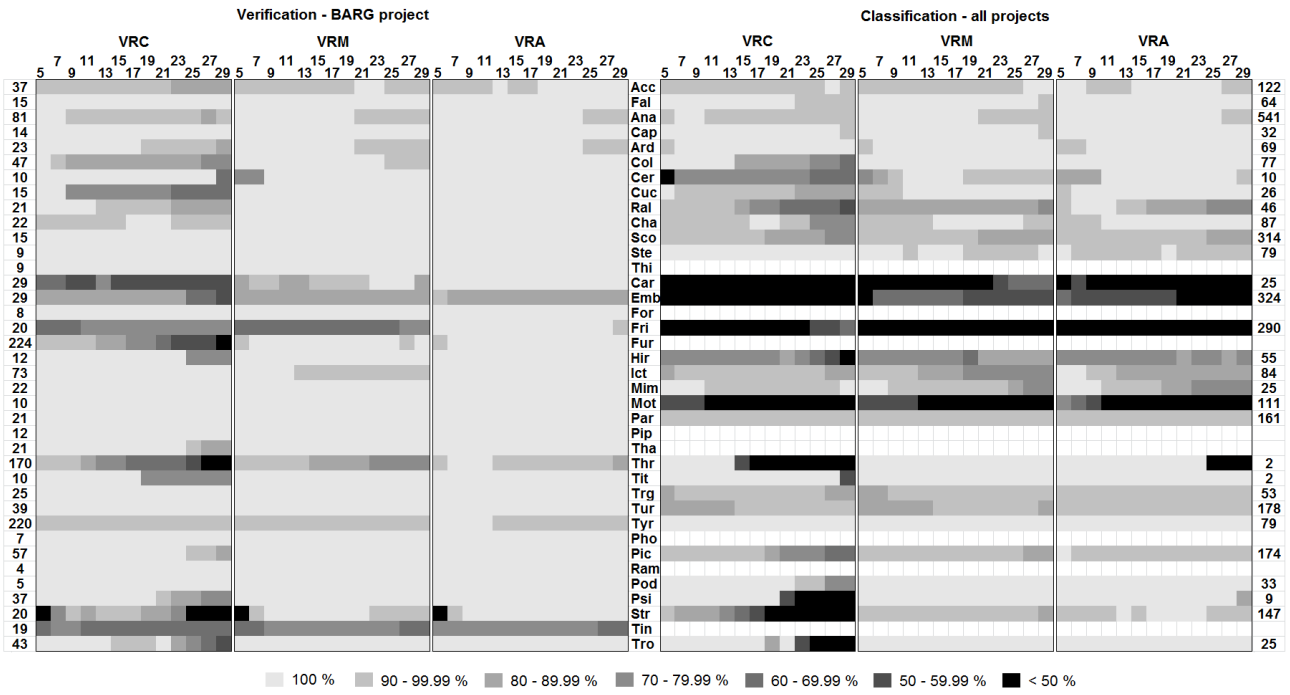


Figure 1. The graphical representation of classification efficiency into bird's families for verification (left) and classification analysis (right) for the VRC, VRM, and VRA method variations.

The left and right numbers are numbers of sequences, top numbers are sizes of moving window, abbreviations in the middle represent family names: **Acc**ipitridae, **Fal**conidae, **Ana**tidae, **Cap**rimulgidae, **Ard**eidae, **Col**umbidae, **Cer**ylidae, **Cuc**ulidae, **Ral**lidae, **Cha**radriidae, **Sco**lopacidae, **Ste**rnidae, **Thi**nocoridae, **Car**dinalidae, **Emb**erizidae, **For**micariidae, **Fri**ngillidae, **Fur**nariidae, **Hir**undinidae, **Ict**eridae, **Mim**idae, **Mot**acillidae, **Par**ulidae, **Pip**ridae, **Tha**mnophilidae, **Thr**aupidae, **Tit**yridae, **Tro**glodytidae, **Tur**didae, **Tyr**annidae, **Pho**enicopteridae, **Pic**idae, **Ram**phastidae, **Pod**icipedidae, **Psi**ttacidae, **Str**igidae, **Tin**amidae, **Tro**chilidae.

Contrary for the VRA method, the family and the order classification efficiencies differ less than 2 %.

Families *Cardinalidae*, *Emberizidae*, *Fringillidae*, and *Thraupidae* are members of *Passeroidea* superfamily subclade nine-primaried oscines. For the VRC, sequences of the first three named families were in most cases classified as *Thraupidae* family. The *Thraupidae* sequences were classified as *Cardinalidae* or *Emberizidae*. The other two method variations were much more successful. There are other members of nine-primaried oscines, *Icteridae* and *Parulidae* which were almost perfectly classified with all method variations.

The classification to *Strigidae* family was in the case of the consensus references very poor; median and average density vectors had significantly better results. Contrary, the classification efficiency for *Tinamidae* family is low for all method variations.

3.2. Classification analysis

A classification analysis was conducted for all of the DNA barcoding projects except BARG. In total, 3244 sequences were classified into 30 families. Some families from the BARG project were not present in the other projects. The right part of the Figure 1 shows graphical representation of the classification efficiencies for each family and size of moving window. The efficiencies are lower than for the BARG sequences because most of the classified sequences belong to different species from different continents than those used for creating the references. As in the case of verification, the VRA and the VRM references are more effective than the VRC references.

The overall results from the Figure 2b confirm better effectiveness of the VRA references for bird classification into orders; the values are above 98 % for all moving window sizes. The efficiencies for family classification are in average 10–15 % lower than efficiencies for orders. The results for the VRM are slightly lower than for the VRA and the VRC showed strong dependence on moving window size.

Sequences of *Passeriformes* order, especially sequences of *Passeroidea* superfamily, were poorly classified into right family but were classified into closely related families like in verification. Classification into *Passeriformes* order was in range of weighted averages 99.71 to 99.93 % for all moving window sizes of the VRA. However, the results for the VRA and VRM are not significantly better than the results of the VRC like in cases of other orders.

4. CONCLUSION

We tested usability of DNA barcodes represented as nucleotide densities for taxonomical classification into bird's orders and families. Our results show that the nucleotide density references created from the consensus sequences for orders and families do not work. The best results were obtained for family references created as average nucleotide densities. The results also show that the size of moving window for ND calculation is not

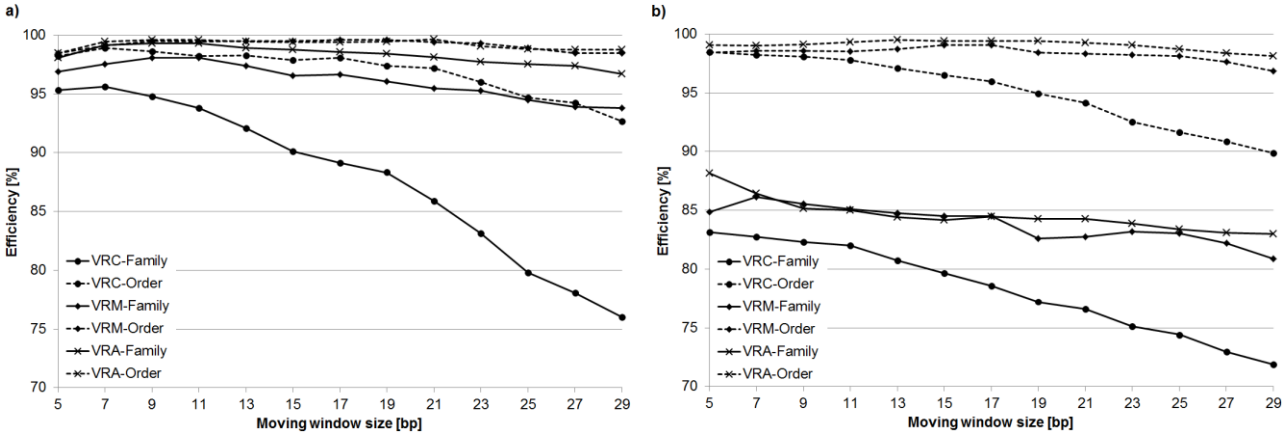


Figure 2. The weighted averages of classification efficiency for all families and orders (a) verification on BARG project, (b) classification of all other projects.

crucial parameter for the VRM and VRA contrary to the VRC.

The obtained results were influenced by accepted morphological taxonomy. The bird's species were sorted according to morphological taxonomy but their molecular relationships of mtDNA may differ. The morphological taxonomy of several species contained in the dataset is still disputed. In particular, the borders between some families of *Passeriformes* order are fuzzy. More than 25 % of the species in the datasets belongs to these families. The reason of low values of family classification efficiencies for *Passeroidea* birds is that this superfamily group has the lowest values of the Euclidean distances between their reference NDs. The average value of the Euclidean distances between references without distances between *Passeriformes* references is 0.0893±0.0122. The average distances between *Passeriformes* references except *Passeroidea* is 0.0914±0.0148. And the distances only between *Passeroidea* is 0.0571±0.0128.

As the proposed method counts Euclidean distances between query sequence and reference sequences we can estimate classification confidence. For example, if the smallest distances belong to not closely related families then the classification is less valid.

Although the family reference databases were created from sequences of South American species, they can also be used with high effectiveness for species originated in other continents. Furthermore, shorter sequences of 600 to 650 bp were not significantly worse classified.

We also conducted a trial identification of species based on the average nucleotide density references. DNA barcoding project ROM-Bats of Guyana 1 (BCBNC) was chosen for calculating references. This project contains 819 specimens of 94 species of barcodes longer than 600 bp. The references were created only for species with minimally 3 specimens. For identification analysis, DNA barcoding project ROM-Bats of Guyana 2 (BCDR) was chosen. The project contains 3779 specimens of barcodes longer than 600 bp. 3594 specimens correspond to the reference species. Our method correctly identified 100.0 % of specimens to corresponding species. The remaining 185 specimens are without corresponding species references. 137 of them were classified to corresponding genus reference, 13 had genus reference but were classified into another genus of the same family, 18 did not have genus reference and were classified into correct family, and the last 17 specimens did not have even family reference.

Recently published work [13] used fuzzy-set-theory approach for species identification based on barcodes. The method was also tested on BCBNC project data, the achieved success rate was 98.2 %. However, the group used the same data to create reference set and testing set in contrary to our standard verification.

Our novel approach to identification and classification based on average nucleotide density references and minimal Euclidean distance between the references and analyzed sequence is promising based on presented classification rate. The method is computationally simple and suitable for fast parallel processing when comparing unknown query sequence with the references. The use of squared Euclidean distance will be also tested.

One of the limiting parameters of all classification or identification comparative methods is validity of taxonomical classification of data in the databases like BOLD or GenBank. Can we be sure that each barcoded specimen was unambiguously identified by experienced taxonomist? And a question above all: Does the barcoding sequence COX1 alone carry enough information for reliable species identification and phylogeny analysis? [14]

From the other hand, the nucleotide density vectors reveal characteristic patterns in sequence. Visual analysis of the barcode's NDs revealed highly conserved regions. We plan to compare conserved regions of homologues sequences with corresponding protein active sites. Furthermore, the correlation of nucleotide density vectors can serve as quick alignment method.

5. ACKNOWLEDGEMENTS

This study was supported by research grant from Czech Science Foundation (No. P102/11/1068) and by European Regional Development Fund – Project FNUSA-ICRC (No. CZ.1.05/1.1.00/02.0123).

6. REFERENCES

[1] C. Mora, D. P. Tittensor, S. Adl, A. G. B. Simpson, and B. Worm, "How many species are there on earth and in the ocean," *PLoS Biology*, vol. 9, e1001127, Aug. 2011.

[2] J. C. Avise, J. Arnold, R. M. Ball, E. Bermingham, T. Lamb, J. E. Neigel, C. A. Reeb, and N. C. Saunders, "Intraspecific phylogeography: the mitochondrial DNA bridge between population genetics and systematics," *Ann. Rev. Ecol. Syst.*, vol. 18, pp. 489–522, 1987.

[3] C. Moritz, T. E. Dowling, and W. M. Brown, "Evolution of animal mitochondrial DNA: relevance for population biology and systematics," *Ann. Rev. Ecol. Syst.*, vol. 18, pp. 269–292, 1987.

[4] D. J. White, J. N. Wolff, M. Pierson, and N. J. Gemmell, "Revealing the hidden complexities of mtDNA inheritance," *Mol. Ecol.*, vol. 17, pp. 4925–4942, Dec. 2008.

[5] M. Kimura, *The neutral theory of molecular evolution*, Cambridge University Press, New York, 1983.

[6] W. M. Brown, M. George, Jr., and A. C. Wilson, "Rapid evolution of animal mitochondrial DNA," *Proc. Natl. Acad. Sci. USA*, vol. 76, pp. 1967 – 1971, Apr. 1979.

[7] N. Galtier, B. Nabholz, S. Glémin, and G. D. Hurst, "Mitochondrial DNA as a marker of molecular diversity: a reappraisal," *Mol. Ecol.*, vol. 18, pp. 4541–4550, Nov. 2009.

[8] P. D. N. Hebert, "Biological identifications through DNA barcodes," *Proc. R. Soc. B: Biological Sciences*, vol. 270, pp. 313 – 321, Feb. 2003.

[9] C. Moritz and C. Cicero, "DNA barcoding: promise and pitfalls," *PLoS Biology*, vol. 2, pp. 1529–1531, Sep. 2004.

[10] H. R. Taylor and W. E. Harris, "An emergent science on the brink of irrelevance: a review of the past 8 years of DNA barcoding," *Mol. Ecol. Resour.*, vol. 12, pp. 377 – 388, Feb. 2012.

[11] F. Austerlitz, O. David, B. Schaeffer, K. Bleakley, M. Olteanu, R. Leblois, M. Veuille, and C. Laredo, "DNA barcode analysis: a comparison of phylogenetic and statistical classification methods," *BMC Bioinformatics*, vol. 10 (Suppl. 14), S10, Nov 2009.

[12] A. L. Bazinet and M. P. Cummings, "A comparative evaluation of sequence classification programs," BMC Bioinformatics, vol. 13:92, 2012.

[13] A.-B. Zhang, C. Muster, H.-B. Liang, C.-D. Zhu, R. Crozier, P. Wan, J. Feng, and R. D. Ward, "A fuzzy-set-theory-based approach to analyse species membership in DNA barcoding," *Mol. Ecol.*, vol. 21, pp. 1848 – 1863, Apr. 2012.

[14] R. DeSalle, "Species Discovery versus Species Identification in DNA Barcoding Efforts: Response to Rubinoff," *Conservation Biology*, vol. 20 (5), pp. 1545-1547, Jun. 2006.

MODIFICATION OF THE *ESCHERICHIA COLI* METABOLIC MODEL IAF1260 BASED ON ANAEROBIC EXPERIMENTS

Jenni J. Seppälä[1], Antti Larjo[2,3], Tommi Aho[1], Anniina Kivistö[1], Matti T. Karp[1] and Ville Santala[1]

[1]Department of Chemistry and Bioengineering, Tampere University of Technology, P.O. Box 541, FI-33101 Tampere, Finland
[2]Department of Signal Processing, Tampere University of Technology, P.O. Box 553, FI-33101 Tampere, Finland[1]
[3]Department of Information and Computer Science, Aalto University School of Science and Technology, Helsinki, Finland
jenni.seppala@tut.fi, antti.larjo@tut.fi, tommi.aho@tut.fi, anniina.kivisto@tut.fi, matti.karp@tut.fi, ville.santala@tut.fi

ABSTRACT

Facultative anaerobic *Escherichia coli* utilize glucose by mixed acid fermentation in the absence of oxygen (O_2). Comprehensive models of *E. coli* have been constructed to predict cellular metabolism. Here, the consistence of predictions with existing metabolic model iAF1260 and experimental data of anaerobic metabolism of *E. coli* is examined.

Experimental data included anaerobic batch experiments with wild type and 30 single gene deletion mutants. Flux balance analysis was applied to iAF1260 to predict the end product distribution and biomass formation of wild type and mutants. Based on the comparison of model simulations and experimental data, we suggest several modifications to the model.

In anaerobic conditions with glucose as substrate, the modified model predicts biomass production that is consistent with the experimental data. The results support the use of modified model for engineering applications of anaerobic metabolism of *E. coli.*

1. INTRODUCTION

E. coli is used as model organism for studying many biological processes. It has useful properties such as rapid growth rate, simple nutritional requirements, well-established genetic tools and a completely sequenced genome. In addition, comprehensive metabolic models have been developed to analyze the metabolism of *E. coli* [1].

E. coli is a facultative anaerobe. In anaerobic conditions it ferments glucose (glc) through mixed acid fermentation, excreting mainly ethanol, acetate, lactate, succinate, carbon dioxide (CO_2) and hydrogen (H_2). The division of the end products is dependent on the given substrate, growth phase and environmental conditions, such as pH [2]. For instance, the optimal NADH/NAD^+ balance and ATP production is maintained by the distribution of fluxes for distinct pathways.

Flux balance analysis (FBA) is a mathematical approach for analyzing the flow of metabolites through a metabolic network [3]. The suitability of existing metabolic reconstruction for the prediction of anaerobic metabolism has been shown, but the essentiality of the genes in the aerobic conditions has been under more exhaustive study than in anaerobic ones [4, 5].

Here we examine the simulation results of anaerobic growth of wild type *E. coli* and its single-gene deletion mutants when glucose is given as substrate. The predictions by model are compared to our previously publish experimental data [6, 7] and modifications improving the prediction accuracy are suggested.

2. MATERIALS AND METHODS

The results of two separate batch experiments were used to examine and modify the metabolic model. The modifications had two aims: first, to modify model to have similar response to single gene knock out as experimentally noticed and, second, to improve the prediction of the end product formation of mixed acid fermentation.

The laboratory experiments are described previously in detail [6, 7] and summarized here.

2.1. Batch experiments with *E. coli*

The first batch experiment had two parallel cultivations of *E. coli* K-12 MG1655 [6]. The cells were cultivated anaerobically in 2100 ml vessels with 250 ml of liquid media. The media had 3 g/l glucose as substrate (full composition is described in [6]). Cultivations were maintained at 37°C for 25.5 hours and samples were taken at 1.5 h interval. From liquid samples acetate, lactate, ethanol, glucose, and biomass (optical density at 600 nm, OD_{600}) were analyzed. H_2 and CO_2 concentrations were measured from gas samples.

For the comparison of the experimental and simulated data, following estimations were made

1. Value 1 at OD_{600} = 0.3 g/l dry cell weight (gDW) [8]

2. Rates of byproduct secretion, biomass production and glucose consumption during two separate time frames, exponential growth phase (3-8h) and stationary growth phase (12-25.5h) were calculated (Figure 1).

In the second batch experiments *E. coli* K-12 BW25113 and its individual gene knockouts were tested [7]. Here we use the results of 30 individual knock outs [5] that are related either to glycolysis or mixed acid fermentation. Deleted genes are listed in Table 1 and shown in Figure 3. Two replicates of wild type and mutants were cultivated and the average results are shown in Table 1. Fermentations were conducted in M9-CA medium (M9 + 1% (w:v) casamino acids) with 3g/l glucose in 27.5 ml anaerobic tubes with 10 ml of liquid media. Cells were maintained over two nights at 37°C with 120 rpm rotation, and then the concentrations of H_2 and CO_2 in the head space were analyzed. Biomass was analyzed by measuring OD_{600}. From six samples the acetate, lactate, ethanol and glucose concentrations were analyzed as in [9], but in room temperature (Figure 2).

2.2. Computational analysis

Cobra toolbox 2.0.5 was used with Matlab R2011b for quantitative prediction of cellular behavior using a constraint-based approach and to visualize the results [10, 11].

2.2.1. Metabolic model

We used the previously constructed genome-scale metabolic network model of *E. coli*, iAF1260, consisting of 1260 genes and containing 1039 metabolites and 2077 reactions [1]. The minimal media composition was used as given in the model with glucose as carbon source. Following modifications were made to the model Ec_iAF1260_flux1 (see corresponding alphabets in Figure 3). Reasoning for the modifications is explained in Section 3.

A) The lower bound of reactions R_EX_o2(e) was set to zero
B) The upper bound of reaction R_FHL changed from zero to unrestricted (=99999)
C) The lower bound of R_EX_co2(e) was set to zero
D) The upper bound of R_LDH_D was set to -1 and gene b2133 (*dld*) was removed from this reaction
E) Reaction R_ACKr was changed to irreversible and genes b3115 (*tdcD*) and b1849 (*purT*) were removed from this reaction.
F) Gene EutD (b2458) was removed from reaction R_PTAr.
G) Gene *mhpF* (b0315) was removed form reaction R_ACALD
H) Reaction R_THRA2i was removed
I) Reaction R_ F6PA was removed.
J) Reactions R_G3PD6 and R_G3PD7 were changed to reversible and lower bounds were set to -1.
K) Reaction R_ME1 was changed to reversible and the lower bound was set to -10.
L) Reaction R_ICDHyr was changed irreversible.
M) Reaction R_AKGDH was removed.

2.2.2. Flux balance analysis

Flux balance analysis (FBA) is a widely used for studying genome scale metabolic networks and simulation of the metabolic capabilities of an organism [12, 3]. The aim is to find a flux distribution **v**, consisting of the reaction rates of all the *n* reactions, under the following constrain:

$$\text{steady-state: } \frac{dx}{dt} = S\boldsymbol{v} = \mathbf{0}, \quad (1)$$

where **x** is the vector of metabolite concentrations and *S* is the stoichiometric matrix. In addition, reaction directionality constrains, medium constraints, and physiological constraints are set

$$\boldsymbol{a} \leq \boldsymbol{v} \leq \boldsymbol{b}, \quad (2)$$

where $\boldsymbol{a}$ and $\boldsymbol{b}$ are vectors containing the lower bound and upper bound for reaction rates in each flux.

The reactions are divided into internal reactions and exchange reactions. Exchange reactions are used for defining the growth medium. Since usually the system is under-determined, an optimization criterion is set on the fluxes in order to get a solution for $\boldsymbol{v}$. This is stated as

$$\min_{\boldsymbol{v}} \boldsymbol{c}^T\boldsymbol{v} \quad (3)$$

where $\boldsymbol{c}$ is a vector determining the objective function as a linear combination of reaction rates. This optimization problem can be solved by linear programming.

The usual assumption in FBA is that the organism tries to maximize its biomass production. That was used as an optimization criterion in this study, too. To visualize the results of FBA, byproduct secretion envelopes (BSE) were calculated for each deletion.

3. RESULTS AND DISCUSSION

The results of the metabolite analysis of the first batch experiment are shown in the Figure 1. The reaction rates derived from these plots are presented along with the BSEs in Figure 3.

Measured endpoint values of the second batch experiment are shown in Table 1 (column C). Additionally, results gained with modified (column B) and unmodified (column A) model are listed. Byproduct secretion by five mutants and wild type was measured and compared with predicted data to ensure total depletion of glucose (Figure 2). Additionally, the exact glucose content of media was measured and calculations of the production per mol of glucose were based on this value. The comparison of the two batch experiment types reveals that in the smaller scale batch cultivation with wild type *E. coli* the ratio of lactate production is higher compared

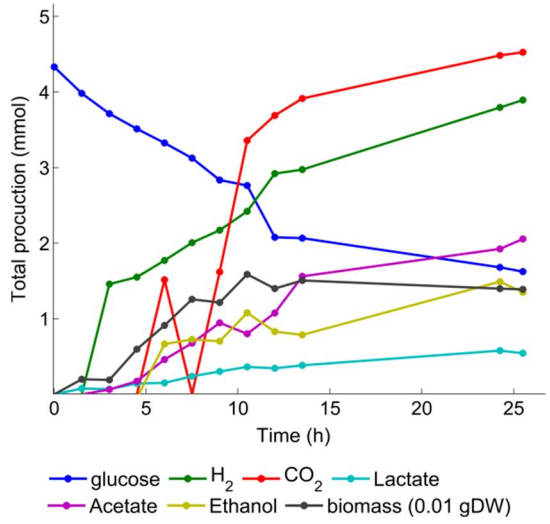


Figure 1. Total byproduct secretion, biomass production and glucose consumption during the first batch experiment. More detailed figures in [6].

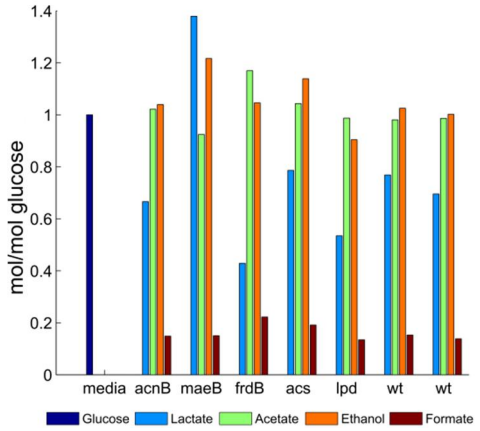


Figure 2. Byproduct secretion results of wild type *E. coli* and five mutants. The results are divided with the amount of glucose consumed (mol/mol glucose), thus the amount of glucose in the zero tube (media) scales to 1.

Deletion		Biomass (gDW/mol glucose)			H_2 production (mol/mol glucose)		
bname	name	A	B	C	A	B	C
WT	WT	21.1	24.7	27.4	1.77	1.57	1.58
b2415	*ptsH*	18.7	22.1	19.4	1.79	1.60	1.64
b2416	*ptsI*	18.7	22.1	1.0	1.79	1.60	0.00
b2943	*galP*	21.1	24.7	18.0	1.77	1.57	1.52
b4025	*pgi*	18.7	22.1	0.3	1.79	1.60	0.00
b3919	*tpiA*	12.3	14.2	22.3	1.86	1.70	1.43
b2463	*maeB*	21.1	24.1	27.5	1.77	1.62	1.71
b0116	*lpd*	21.1	24.7	27.5	1.77	1.57	1.61
b0115	*aceF*	21.1	24.7	20.5	1.77	1.57	1.55
b0114	*aceE*	21.1	24.7	20.2	1.77	1.57	1.55
b2976	*glcB*	21.1	24.7	27.2	1.77	1.57	1.56
b4069	*acs*	21.1	24.7	25.9	1.77	1.57	1.51
b1380	*ldhA*	21.1	25.6	17.7	1.77	1.68	1.46
b1478	*adhP*	21.1	24.7	22.6	1.77	1.57	1.45
b2224	*atoB*	21.1	24.7	21.9	1.77	1.57	1.41
b0871	*poxB*	21.1	24.7	20.7	1.77	1.57	1.40
b3956	*ppc*	0.0	24.7	21.1	2.55	1.57	1.34
b2296	*ackA*	21.1	18.3	20.6	1.77	0.10	0.12
b1702	*ppsA*	21.1	24.7	7.9	1.77	1.57	0.08
b2297	*pta*	21.1	18.3	15.9	1.77	0.10	0.00
b1241	*adhE*	21.1	22.9	13.7	1.77	0.92	0.15
b3403	*pck*	21.1	24.7	21.9	1.77	1.57	1.54
b0118	*acnB*	21.1	24.7	18.8	1.77	1.57	1.71
b3236	*mdh*	21.0	23.9	22.0	1.77	1.62	1.71
b4153	*frdB*	0.0	24.7	23.7	0.00	1.57	1.81
b4152	*frdC*	0.0	24.7	20.8	0.00	1.57	1.71
b4154	*frdA*	0.0	24.7	17.0	0.00	1.57	1.65
b4151	*frdD*	0.0	24.7	19.5	0.00	1.57	1.70
b4122	*fumB*	21.1	24.7	18.9	1.77	1.57	1.63
b1611	*fumC*	21.1	24.7	19.6	1.77	1.57	1.60
b1612	*fumA*	21.1	24.7	23.1	1.77	1.57	1.59

Table 1. The predicted and experimental reasults of biomass and H_2 production per mol of glucose. The colums are A) predicted results with original model. B) predicted results with modified model and C) experimental results

to other metabolites than within larger batch cultivation. Nevertheless, all the cultures produce lactate. In the gene deletion experiments, all tested mutants were able to grow in the minimal media (Table 1).

The metabolic model of *E. coli* was altered based on batch experiments. The main criteria for the modifications were:

1. Maintaining the measured values of the first batch experiment with in the BSEs
2. Enhancing the consistency of the simulation results and experimental data of *E. coli* mutants. Mainly mutations causing major difference in predicted and experimental biomass production were inspected.

In the following, the motivations and the effects of changes made to the model will be described. The two first modifications, A and B, were applied in every simulation.

A) The uptake of oxygen was closed to make the metabolism anaerobic.
B) In the original model the reaction producing H_2 and CO_2 from formate was closed to prevent the aerobic production of H_2. Here this pathway was opened to allow the H_2 production.
C) The uptake of CO_2 was set to zero, since *E. coli* is not assumed to assimilate CO_2 [13].
D) The model was forced to produce at least 0.125 mol of lactate per mol of glucose. The unmodified model does not produce lactate during the maximum growth. This is due to reactions of the mixed acid fermentation, where the production of lactate from pyruvate (pyr) produces only one NAD^+ whereas, *e.g.* the production of ethanol from pyruvate produc-

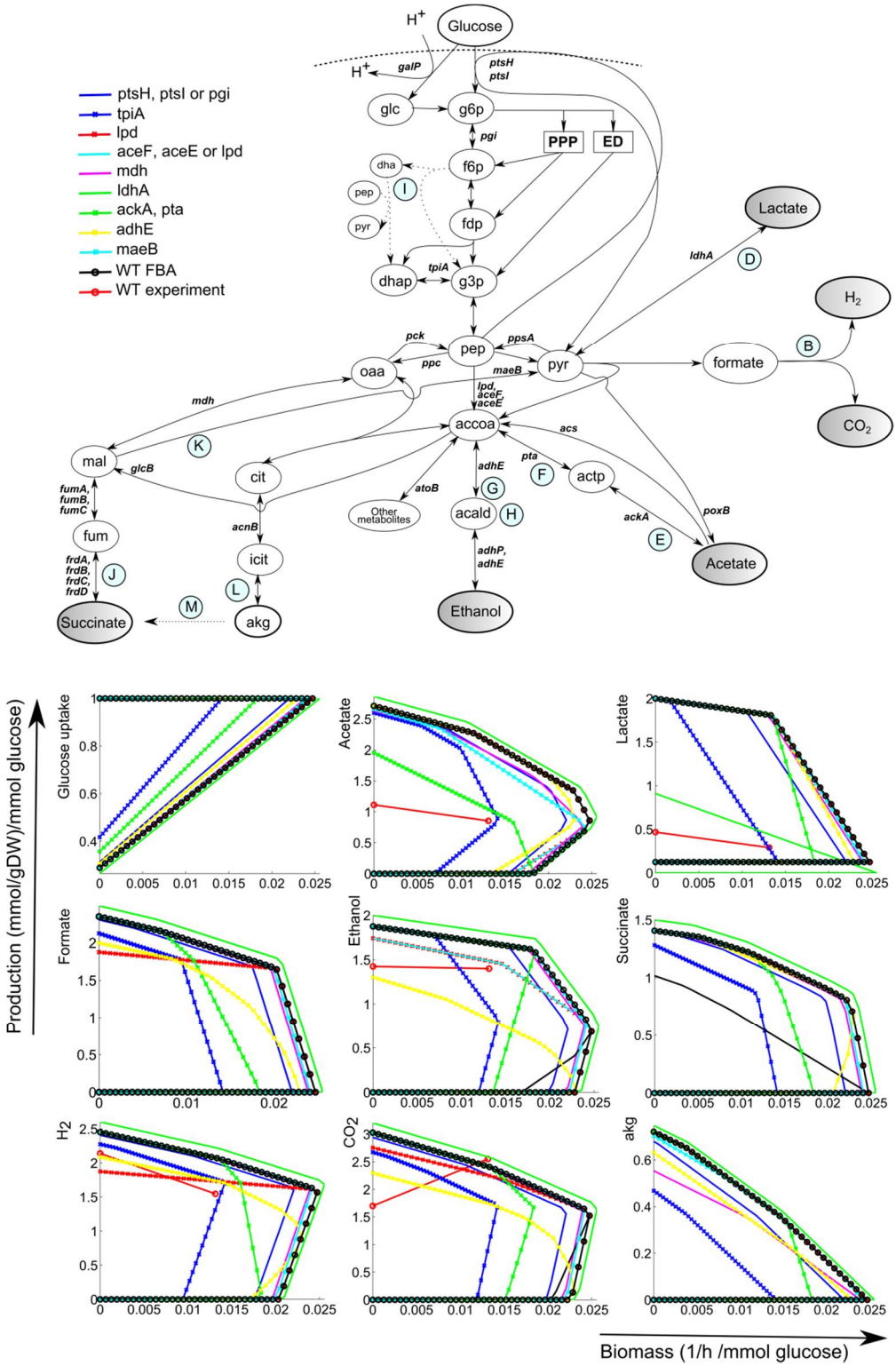


Figure 3. Above, metabolic pathway of glycolysis and mixed acid fermentation, conducted based on metabolic model and flux distributions with FBA. Circled letters refer to the modifications describer within the text. Below, byproduct secretion envelopes (BSE) predicted with the modified model. BSEs present the space of possible solutions to byproduct secretion with given biomass. In this case, all the BSEs limit to a single value when the biomass is maximized. The experimental results are obtained from the data shown in the Figure 1. (PPP = pentose phosphate pathway, ED = Entner–Doudoroff pathway, fdp = fructose-1,6-bisphosphate, g6p = α-D-glucose 6-phosphate, fum = fumarate, cit = citrate)

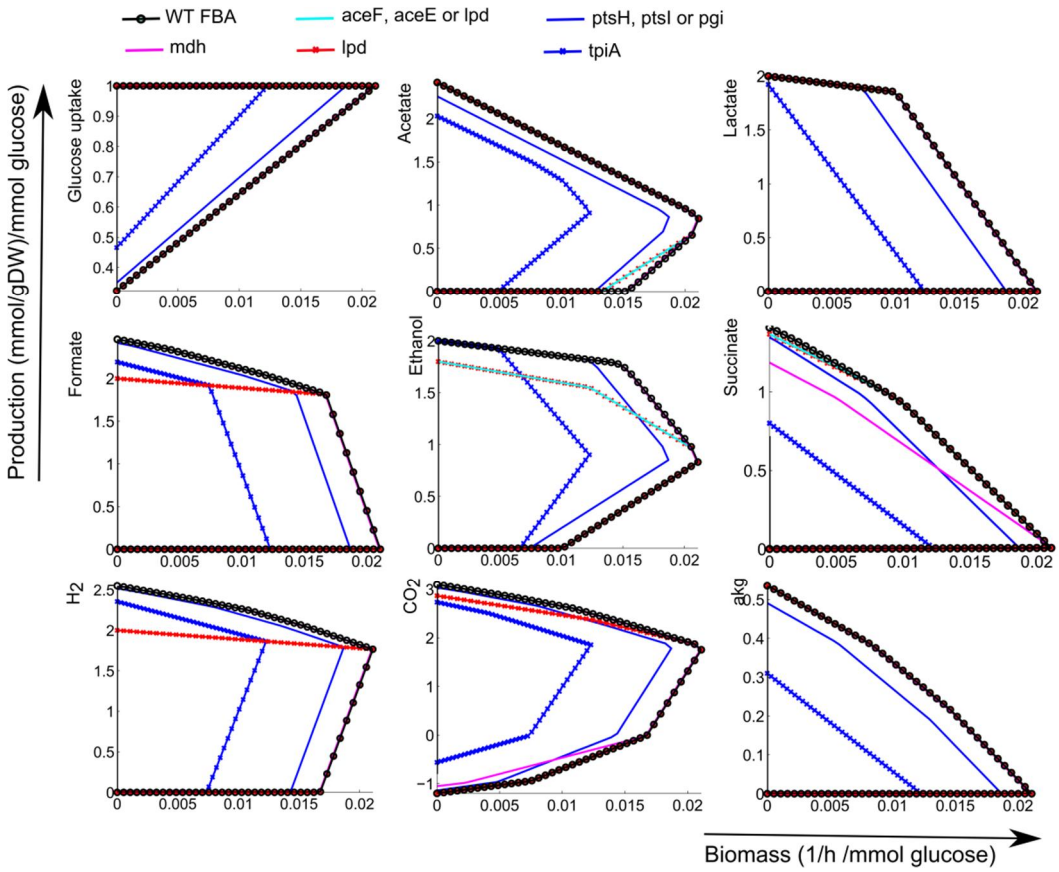


Figure 4. Byproduct secretion envelopes based FBA of unmodified metabolic model.

es two NAD^+s and acetaldehyde (acald). For this reason, simulations imply that *E. coli* is able to produce more biomass if it does not produce lactate. However, in the batch experiments, all the measured samples contained lactate (Figure 2) therefore the forced production of lactate was added to the model. It has been shown that the activity of lactate dehydrogenase (LDH) is increased 10-fold at low pH [14]. It is known that variation in pH causes changes in the end product distribution, which cannot be described by stoichiometric model [2]. In batch experiments the pH is not controlled, thus it decreases over the experiment as shown in [6].

In the model, the production of D-lactate from pyruvate was catalyzed by two enzymes: LdhA and Dld. *Dld* was removed, since it is mainly required for aerobic growth on lactate [13]. This removal caused the simulated *ldhA* mutant to have decrease in lactate production and increase in H_2 production based on glucose consumed. These effects have been detected in previous studies [15]. In addition, along with removal of forced lactate production (*ldhA*), the predicted maximum biomass is increased (Figure 3).

E) The reaction ACKr is reversible production of acetylphosphate (actp) from acetate. The reaction was changed to irreversible such that the production of acetylphosphate from acetate was blocked. It is thought that if *E. coli* has glucose to consume, it first consumes all the glucose before switching to acetate metabolism [16]. In the model, the modified reaction can be independently catalyzed by TdcD (b3115), AckA (b2296) or PurT (b1849). The genes *tdcD* and *purT* were removed, causing the prediction of the deletion of *ackA* to have decrease in the H_2 production as seen in the experiments. Overexpression of *tdcD* has been found to partially repair the growth effect of the deletion of *ackA* [17], but the gene is mainly involved in propanoate metabolism. The acetate production catalyzed by *purT* is a side reaction in the purine pathway, thus not affecting directly to mixed acid fermentation. In *E. coli* the acetylphosphate formation via the products of *pta* and *ackA* is assumed to be the only source of this key metabolite, which is involved in many cellular processes [18].

F) In order to improve the prediction accuracy in the case of *pta* deletion, *eutD* (b2458) was removed from the reversible reaction forming acetylphosphate from acetyl-coA (accoa). Previously it has been shown, that the overexpression of *eutD* gene can compensate for the double deletion of *acs* and *pta* when grown aerobically on acetate [16, 19], but here glucose is used as substrate. Additionally, the experimental gene deletion data support the removal of *eutD* from this reaction (Table 1).

G) According to the model, the reversible reaction producing acetaldehyde (acald) from acetyl-coA (accoa) can be independently catalyzed by either MhpF or AdhE. *MhpF* was removed from this reaction, since it was experimentally observed that the deletion of *adhE* influences the growth and biomass production (Table 1). It has been proposed that mhpF might catalyze this reaction, but its properties have not been biochemically characterized [20].

H) The removed reaction THRA2i is irreversible transformation of L-allo-threonine to glycine and acetaldehyde, catalyzed alternatively by GlyA or LtaE. FBA suggest that this reaction enables an alternative pathway for the production of acetaldehyde under the *adhE* gene deletion. When this pathway was removed, some decrease in the H_2 and biomass production occurred in the predictions with the deletion of *adhE*, which was also experimentally detected.

I) The removal of the reaction degrading fructose-6-phosphate (f6p) to glycerone (dha) and glyseraldehyde-3-phosphate (g3p) from the model did not change the end products of the fermentation. However, it directed the fluxes through generally accepted routs of the glycolysis. The two genes designated in the model for this reaction, *fsaA* and *fsaB*, are found from *E. coli*, but their physiological role in cell is not yet clear [21]. In the model, reactions DHAPT and F6PA together form an alternative route for glucose degradation, designated with I in Figure 3. In literature, this route is not assumed to be in use, but FBA often direct the fluxes through this pathway. It has no effect to the predicted biomass production.

J) The deletions of fumarate reductase (*frdA, frdB, frdC and frdD*) were experimentally non-lethal (Table 1). However, the predicted effect is lethal in the original model. In order to correct this, reactions G3PD6 and G3PD7 were changed reversible with restricted flows. The reactions included the conversion of glycerol-3-phosphate (g3p) to dihydroxyacetone (dhap) which is catalyzed by anaerobic glycerol-3-phosphate-dehydrogenase and related to phospholipid metabolism. These reactions are reversible in the EcoCyc database [22]. The modification was implemented in order to compensate the predicted production blockage of menaquinol-8 (mql8) and demethylmenaquinol-8 (2dmmql8), caused by deletion of fumarate reductase. With these changes the fumarate reductase mutant was predicted have the same maximum biomass production rate as the wild type.

K) The original model predicted the deletion of *ppc* to be lethal, but this was not the case with the experimental data (Table 1). To enable the growth of the *ppc* mutant, the production of pyruvate from malate (mal) catalyzed by MaeA, was changed to reversible. Malic enzymes can catalyze the reaction also for physiologically non-favored direction [23].

L) The reversible production of 2-oxoglutarate (akg) from isocitrate (icit), catalyzed by isocitrate dehydrogenase, was changed to be an irreversible reaction working to the physiologically favored direction [22]. This was done to prevent the glutamate utilization via this reaction, which occurred in our preliminary calculations (rich media composition including glutamate). Here, glucose as the sole carbon source, this change has no effect.

M) To ensure that the citric acid cycle is not used, the pathway from akg to succinyl-coA was removed. These reactions are assumed to be repressed in anaerobic conditions [24].

Before model modification, 17 of the 30 mutants were predicted to have identical metabolite production as wild type, 5 not to grow and 8 to have changes in the metabolite production (Figure 4). After the modifications 17 of the mutants were identical with wild type and 13 mutants were predicted to have effect to the metabolite production. All mutants were predicted to grow in the minimal media. The model modifications were able to correct all but one of the major mispredictions of the original model. The mutant with deletion of *ppsA*, which gene product takes part to a reaction where pyruvate is used to form phosphoenolpyruvate (pep), is still predicted to be identical with wild type, even though in experiments the deletion had severe effect on the growth.

The experimental data fits within the predicted byproduct secretion envelopes. This supports the validity of the modified model for the simulation of anaerobic metabolism of *E. coli.*

The modified stoichiometric model includes a large set of reactions and genes. The modifications here include removals of reactions and genes, and changes in the directions of reactions. Neither reactions nor genes were added. The underlying comprehensive model includes all known reactions, and those can be modified based on own data and experimental conditions to give more accurate estimations for a specific requirements.

4. CONCLUSIONS

Experimental data related to mixed acid fermentation was compared with model predictions. A previously presented metabolic model was refined in order to unify the simulated and experimental results. Here, the suggested modifications were described and the simulation results of the original and modified model were presented. Modifications were shown to improve the prediction of the essentiality of the studied genes. The essentiality predictions and experimental results were shown to be coherent, thus the model can be used for the engineering applications of anaerobic metabolism of *E. coli.*

5. ACKNOWLEDGEMENTS

This research was funded by the Academy of Finland (project no 126974 and 140018) and Tampere University of Technology Graduate School (J. Seppälä). The work was also supported by the Academy of Finland (application number 213462, Finnish Programme for Center of Excellence in Research 2006-2011).

6. REFERENCES

[1] Feist AM, Henry CS, Reed JL, Krummenacker M, Joyce AR, Karp PD, Broadbelt LJ, Hatzimanikatis V and Palsson BØ, "A genome-scale metabolic reconstruction for *Escherichia coli* K-12 MG1655 that accounts for 1260 ORFs and thermodynamic information," *Molecular Systems Biology*, 3:121, 2007

[2] N. Mnatsakanyan, K. Bagramyan and A. Trchounian, "Hydrogenase 3 but not hydrogenase 4 is major in hydrogen gas production by *Escherichia coli* formate hydrogen-lyase at acidic pH and in the presence of external formate," *Cell Biochemistry and Biophysics*, vol. 41, pp. 357-365, November 2004

[3] J. D. Orth, I. Thiele, and B. Ø. Palsson, "What is flux balance analysis?" *Nature Biotechnology*, vol.28, pp. 245–248, March 2010

[4] J. S. Edwards, R.U. Ibarra and B. Ø. Palsson, "In silico predictions of *Escherichia coli* metabolic capabilities are consistent with experimental data," *Nature Biotechnology*, vol. 19, pp. 125–130, 2001.

[5] T. Baba, T. Ara, M. Hasegawa, Y. Takai, Y. Okumura, M. Baba, K. A. Datsenko, M. Tomita, B. L. Wanner, H. Mori, "Construction of *Escherichia coli* K-12 in-frame, single-gene knockout mutants: the Keio collection," *Molecular Systems Biology*, 2:2006.0008, February 2006.

[6] J. J. Seppälä, J. A. Puhakka, O. Yli-Harja, M. T. Karp, V. Santala, "Fermentative hydrogen production by *Clostridium butyricum* and *Escherichia coli* in pure and cocultures," *International Journal of Hydrogen Energy*, vol. 36, pp. 10701-10708, August 2011.

[7] J.J. Seppälä, A. Larjo, T. Aho, O.Yli-Harja, M.T. Karp, V. Santala, "Prospecting hydrogen production of Escherichia coli by metabolic network modeling" (Submitted)

[8] J. Soini, K. Ukkonen and P. Neubauer, "High cell density media for Escherichia coli are generally designed for aerobic cultivations – consequences for large-scale bioprocesses and shake flask cultures," *Microbial Cell Factories* 7:26, August 2008.

[9] A. Kivistö, V. Santala and M. Karp, "Hydrogen production from glycerol using halophilic fermentative bacteria," *Bioresource Technology*, vol. 101 pp. 8671-7, November 2010.

[10] S. A. Becker, A. M. Feist, M. L. Mo, G. Hannum, B. Ø. Palsson and M. J. Herrgard, "Quantitative prediction of cellular metabolism with constraint-based models: The COBRA Toolbox", *Nature Protocols*, vol. 2, pp. 727-738, 2007.

[11] Mathworks Inc., "Matlab R2011b", http://mathworks.com/

[12] D. A. Fell and J.R. Small, "Fat synthesis in adipose tissue. An examination of stoichiometric constraints," *Biochemical Journal*, vol. 238, pp. 781-786, 1986.

[13] M. R. Parikh, D. N. Greene, K. K. Woods, and I. Matsumura, "Directed evolution of RuBisCO hypermorphs through genetic selection in engineered *E. coli," Protein Engineering, Design and Selection*, vol. 19, pp. 113-119, March 2006

[14] P. K. Bunch, F. Mat-Jan, N. Lee, and D. P. Clark, "The *IdhA* Gene Encoding the fermentative lactate dehydrogenase of *Escherichia coli", Microbiology*, vol. 143, pp. 187-195, January 1997.

[15] A. Yoshida, T. Nishimura, H. Kawaguchi, M. Inui, H. Yukawa, "Enhanced hydrogen production from glucose using ldh- and frd-inactivated *Escherichia coli* strains," *Applied Microbiology and Biotechnology*, vol. 73, pp. 67-72, November 2006.

[16] A. J. Wolfe, "The Acetate Switch," *Microbiology and Molecular Biology Reviews*, vol. 69, pp. 12-50, March 2005.

[17] C. Hesslinger, S. A.Fairhurst, G. Sawers, "Novel keto acid formate-lyase and propionate kinase enzymes are components of an anaerobic pathway in *Escherichia coli* that degrades L-threonine to propionate," *Molecular Microbiology*, vol. 27, pp. 477-492, February 1998.

[18] I. Mizrahi, D. Biran and E. Z. Ron, "Requirement for the acetyl phosphate pathway in *Escherichia coli* ATP-dependent proteolysis," *Molecular Microbiology*, vol. 62, pp. 201-11, 2006.

[19] F. P. Bologna, V. A. Campos-Bermudez, D. D. Saavedra, C. S. Andreo, M. F. Drincovich, "Characterization of *Escherichia coli* EutD: a phosphotransacetylase of the ethanolamine operon ," *The Journal of Microbiology*, vol. 48, pp 629-636, October 2010.

[20] A. Ferrández, J. L. Garciá, E. Díaz, "Genetic characterization and expression in heterologous hosts of the 3-(3-hydroxyphenyl)propionate catabolic pathway of *Escherichia coli* K-12," Journal of Bacteriology, vol. 179, pp. 2573-81, April 1997.

[21] M. Schurmann and G. A. Sprenger, "Fructose-6-phosphate aldolase is a novel class I aldolase from *Escherichia coli* and is related to a novel group of bacterial transaldolases," *The Journal of Biological Chemistry*, vol. 276, pp.11055-61, 2001.

[22] I. M. Keseler, J. Collado-Vides, A. Santos-Zavaleta, M. Peralta-Gil, S. Gama-Castro, L. Muniz-Rascado, C. Bonavides-Martinez, S. Paley, M. Krummenacker, T. Altman et al., "EcoCyc: a comprehensive database of *Escherichia coli* biology,". *Nucleic Acids Research*, 39:D583-D590, January 2011.

[23] L. Stols and M. I. Donnelly "Production of succinic acid through overexpression of NAD(+)-dependent malic enzyme in an *Escherichia coli* mutant," *Applied and Environmental Microbiology*, vol. 63, pp. 2695-701, July 1997.

[24] F. Mat-Jan, C. R. Williams and D. P. Clark, "Anaerobic growth defects resulting from gene fusions affecting succinyl-CoA synthetase in *Escherichia coli* K12," *Molecular and General Genetics*, vol. 215, pp. 276-280, January 1989.

INVERSE MODELING OF THE *DROSOPHILA* GAP GENE SYSTEM: SPARSITY PROMOTING BAYESIAN PARAMETER ESTIMATION AND UNCERTAINTY QUANTIFICATION

Nikolaos Sfakianakis[1] *and Martin Simon*[1]

[1] Department of Mathematics, University of Mainz, Germany

ABSTRACT

In this work we propose a regularization strategy for the inverse parameter identification problem for the *Drosophila* gap gene circuit model in the framework of Bayesian inversion. In contrast to classical deterministic methods the proposed scheme aims not only to find approximate parameter values but also to estimate their reliability.

1. INTRODUCTION

Many biological processes may be modeled by gene regulatory networks involving systems of ordinary differential equations (ODEs) which typically depend on a number of parameters. A major obstacle in the mathematical modeling of gap gene circuits is the fact that these parameters can not be directly assessed by measurement and hence have to be indirectly inferred from experimental data. For the *Drosophila* gap gene system this data usually consists of protein concentrations measured with a certain limited accuracy for a finite set of observation times.

Such *inverse parameter identification problems* are typically ill-posed due to the fact that the available experimental data may not suffice to uniquely determine the parameters. Furthermore, even if there is a unique solution, it may not depend continuously on the data. That is, small measurement errors will lead to large error propagation from the data to the solution and hence way-off parameter estimates. This effect becomes more dominant with increasing number of unknown parameters and for problems with many parameters it is hence necessary to apply some form of regularization. There is a vast mathematical literature on the regularization of inverse and ill-posed problems, for the mathematical theory of deterministic regularization methods see e.g. Engl, Hanke and Neubauer [1] and for statistical inverse problems in the Bayesian framework see e.g. Kaipio and Somersalo [2] and the references therein.

This work is inspired by the paper [3] by Ashyraliev, Jaeger and Blom, who thoroughly analyze the quality of parameter estimates for the *Drosophila* gap gene system obtained by employing the Levenberg-Marquardt method and come to the conclusion that none of the parameters of the full model for the *Drosophila* gap gene system can be determined individually with reasonable accuracy due to extreme correlations between parameters.

Parameter correlation is, however, common in inverse and ill-posed problems, where it results in non-uniqueness of the estimated parameter values. We address this issue by proposing a three-stage numerical method in the framework of Bayesian inversion, combining a priori sensitivity analysis, a sparsity promoting maximum a posteriori (MAP) estimate and subsequent uncertainty quantification using Markov chain Monte Carlo (MCMC) sampling of the posterior density. The basic idea behind this strategy is as follows: The first stage of the algorithm aims at finding a set of indices corresponding to parameters which may be not reliably identifiable from the available data. The second stage then selects among several solutions compatible with the measured data one where the maximum number of parameters from the index set are equal to zero. This yields a reduced gap gene circuit model, nevertheless capable of reproducing the experimental data. Sparsity promoting regularization has been proposed and successfully used in practice by Kuegler et al. [4], who study parameter estimation in chemical reaction systems. See also the survey article [5] by Engl et al. for other biological applications of regularization strategies with sub-linear l^p penalty term. Finally the third stage of the proposed algorithm is designed to asses the reliability of the estimated parameter values. We use MCMC sampling techniques in order to quantify the inherent uncertainty, reflecting both, modeling and measurement errors.

2. THE *DROSOPHILA* GAP GENE NETWORK

In this work we focus on the gap gene system, important in the anterior-posterior (A-P) specification of the vinegar fly *Drosophila melanogaster*, which has been extensively studied, e.g. by Nuesslein-Vollhard, Igham, Akam and collaborators, see [6, 7, 8, 9]. Here, the gap gene system represents one regulatory tier within a hierarchy of mutually regulating genes with the function of forming a molecular pre-pattern of the embryo. Maternal coordinate genes *bicoid* (*bcd*), *hunchback* (*hb*) and *caudal* (*cad*) control expression of the gap genes *giant* (*gt*), *hb*, *Kruppel* (*Kr*) and *knirps* (*kni*), such that these are expressed in one or more broad overlapping domains along the embryos A-P axis. All gap genes constitute transcription factors. The mathematical modeling of the *Drosophila* gap gene system was initiated by Reinitz, Jaeger and collaborators [10, 11, 12, 3]. These authors make the modeling assump-

tion that in the early stages of development, the nuclei of *Drosophila* are arranged in a row, along the A-P axis of the cell. The model they derive describes the changes in the concentrations of the gap gene proteins in every nucleus over time using a corresponding one dimensional model.

The system of ODEs that constitute the model include transcriptional cross-regulation of genes as well as protein production, decay and diffusion, and is given e.g. in [11, 3] as:

$$\begin{aligned}\frac{\mathrm{d}y_i^a}{\mathrm{d}t} = \; & R_a \Phi\left(\sum_{b=1}^{N} W_a^b y_i^b + m_a y_i^{\mathrm{Bcd}} + h_a\right) \\ & -\lambda_a y_i^a + D_a(y_{i+1}^a - 2y_i^a + y_{i-1}^a), \end{aligned} \tag{1}$$

where $y_i^a(t)$ represents concentration of the product of the gene $a = 1, \ldots, N$ in the nucleus $i = 1, \ldots, M$ at time t. R_a are the promoter strengths of the gene a (rate of protein synthesis of a from mRNA). The function Φ is the *regulation expression* function, and is assumed to have the form

$$\Phi(x) = \frac{1}{2}\left(\frac{x}{\sqrt{x^2+1}} + 1\right). \tag{2}$$

W_a^b represents the regulation of gene a by gene b. y_i^{Bcd} denotes the concentration of Bcd protein in nucleus i and m_a the regulation of Bcd on gene a. Moreover h_a, λ_a, and D_a represent the promoter threshold, the decay rate and the diffusion coefficient respectively.

3. FORWARD PROBLEM

The so-called *forward problem* for (1) is well-posed in the sense of Hadamard, i.e. for all admissible data there exists a unique solution that depends continuously on the data. The system (1) consists of $n = N \cdot M$ equations and is governed by a set of $m = N^2 + 5N$ parameters, i.e. $\mathbf{R} \in \mathbb{R}^N$, $\mathbf{W} \in \mathbb{R}^{N\times N}$, $\mathbf{m} \in \mathbb{R}^N$, $\mathbf{h} \in \mathbb{R}^N$, $\lambda \in \mathbb{R}^N$, $\mathbf{D} \in \mathbb{R}^N$. Given these parameters along with, the constant in time $\mathbf{y}^{\mathrm{Bcd}} \in \mathbb{R}^M$, the initial conditions $\mathbf{y}_0 \in \mathbb{R}^n$ of $\mathbf{y}$ at a time $t = 0$, and assuming homogeneous Neumann boundary conditions, we can compute a numerical approximation of the product concentrations y predicted by the system (1). For the sake of readability we "vectorize" the full set of parameters as $\mathbf{x} \in \mathbb{R}^m$ with $\mathbf{x} = [\mathbf{R}, \tilde{\mathbf{W}}, \mathbf{m}, \mathbf{h}, \lambda, \mathbf{D}]^T$, where $\tilde{\mathbf{W}}$ denotes a vector in $\mathbb{R}^{N\cdot N}$ containing the entries of $\mathbf{W}$. We define the discrete forward operator $\mathbf{F}_h(\cdot, t)$ that maps the vector $\mathbf{x} \in \mathbb{R}^m$ onto the numerical approximation $\mathbf{y}_h(t) = [y_h^1(t), \ldots, y_h^n(t)]^T \in (\mathbb{R}_0^+)^n$ of the solution of the system of ODEs (1), where h denotes the smallest timestep used in the numerical discretization scheme. For the numerical computation of $\mathbf{y}_h$ we use the BDF method with Newton iteration from the SUNDIALS CVODES suite.

Finally, assuming independent, additive noise, the measured data corresponding to the unknown parameter vector $\mathbf{x} \in \mathbb{R}^m$ is of the form

$$\mathbf{y}_{\mathrm{m}}(t) = \mathbf{F}_h(\mathbf{x}, t) + \mathbf{e}, \tag{3}$$

where the error term $\mathbf{e}$ includes all kinds of errors involved, i.e. modeling as well as measurement errors.

4. INVERSE PROBLEM

4.1. Bayesian framework

In the Bayesian framework all quantities are modeled as random variables which reflects the uncertainty inherent not only in experimentally measured data but also in mathematical models and their computational implementation. We indicate this by using capital letters for random variables and lower case letters for their particular realizations. The forward model (3) is thus written in the form

$$\mathbf{Y}(t) = \mathbf{F}_h(\mathbf{X}, t) + \mathbf{E}. \tag{4}$$

For simplicity we assume here $\mathbf{E} \sim \mathcal{N}(0, \sigma^2 \mathbf{I}_n)$, $\sigma > 0$, where $\mathbf{I}_n$ is the n-dimensional identity matrix. Now let $\mathbf{Y} := [\mathbf{Y}(t_1)^T, ..., \mathbf{Y}(t_{N_{\mathrm{m}}})^T]^T$ contain measurements at a finite set of observation times. Then the conditional probability density $\pi(\mathbf{y}|\mathbf{x})$ is

$$\frac{1}{(2\pi\sigma^2)^{\frac{n}{2}}} \exp\left(-\frac{1}{2\sigma^2}\sum_{j=1}^{N_{\mathrm{m}}} ||\mathbf{y}(t_j) - \mathbf{F}_h(\mathbf{x}, t_j)||_2^2\right). \tag{5}$$

Assuming that the measurement data $\mathbf{y}_{\mathrm{m}}$ is given, Bayes' theorem yields the *posterior distribution* of $\mathbf{X}$, conditioned on the measured data

$$\pi(\mathbf{x}|\mathbf{y}_{\mathrm{m}}) = \frac{\pi(\mathbf{y}|\mathbf{x})\pi_{\mathrm{pr}}(\mathbf{x})}{\pi(\mathbf{y}_{\mathrm{m}})}, \tag{6}$$

where π_{pr} is the *prior density* encoding all available a priori information about the unknown parameter vector $\mathbf{x}$. From this posterior distribution of the unknown, it is common to calculate various *point estimates*, the most prominent ones being the *maximum a posteriori* (MAP) estimate, given by the highest mode of the posterior

$$\mathbf{x}_{\mathrm{MAP}} := \underset{\mathbf{x}\in\mathbb{R}^m}{\operatorname{argmax}}\{\pi(\mathbf{x}|\mathbf{y}_{\mathrm{m}})\} \tag{7}$$

and the *conditional mean* (CM) estimate, i.e. the expectation of the posterior distribution

$$\mathbf{x}_{\mathrm{CM}} := \mathbb{E}[\mathbf{x}|\mathbf{y}_{\mathrm{m}}] = \int_{\mathbb{R}^m} \mathbf{x}\pi(\mathbf{x}|\mathbf{y}_{\mathrm{m}})d\mathbf{x}. \tag{8}$$

The computation of $\mathbf{x}_{\mathrm{MAP}}$ leads to an optimization problem, whereas computing $\mathbf{x}_{\mathrm{CM}}$ is a high dimensional integration problem.

4.2. Stage 1: A priori sensitivity analysis

In this work we always assume that we already have some initial guess $\mathbf{x}_0$ together with a priori lower and upper bounds for the unknown parameters obtained for instance by a global search algorithm, cf. [3]. In the first stage we compute the sensitivities with respect to the full set of unknown parameters at the initial guess $\mathbf{x}_0$. A parameter is called *non-identifiable* if for given measurement data the objective functional of interest is not sensitive to variations in that parameter, i.e. if the absolute value of the sensitivity is below a certain threshold. If some parameter

x_k turns out to be non-identifiable, its index k is added to an index set I_0. For the numerical implementation of the first stage we use the SUNDIALS CVODES adjoint solver included in the SMBL ODE Solver library, see Lu et al. [13]. For a description of adjoint sensitivity analysis in the case of discrete measurement data we refer the reader to the same reference.

4.3. Stage 2: Sparsity promoting MAP estimate

Let us consider the full set of m unknowns with second stage prior given by the Gibbs distribution

$$\pi_{\text{pr},2}(\mathbf{x}) \propto \exp(-\alpha||\mathbf{x}||_p^p), \tag{9}$$

where $\alpha > 0$ is a regularization parameter and

$$||\mathbf{v}||_p := \left(\sum_{k \in I_0} |v_k|^p \right)^{1/p}, \; p \in (0, 2]. \tag{10}$$

The MAP estimate is thus given by

$$\underset{\mathbf{x}\in\mathbb{R}^m}{\operatorname{argmin}} \{ \sum_{j=1}^{N_\text{m}} ||\mathbf{y}_\text{m}(t_j) - \mathbf{F}_h(\mathbf{x}, t_j)||_2^2 + 2\sigma^2\alpha||\mathbf{x}||_p^p \}. \tag{11}$$

The choice $p = 2$ yields the classical *Tikhonov regularization*, cf. [1], however with regard to the aforementioned non-uniqueness and correlation issues, we are here mainly interested in sparse solutions, that is parameter vectors $\mathbf{x}$ with $x_k = 0$ for as many $k \in I_0$ as possible. It is thus natural to consider more general regularization strategies with some $p \in (0, 2)$, which is motivated by the observation that for decreasing value of p the sparsity enforcement becomes stronger. Notice moreover that despite the fact that for $p < 1$ the penalty functional is no longer convex, it has been shown by Grasmaier [14] that also for these values of p the estimate (11) yields a regularization strategy. As described in the reference [4], we solve the optimization problem numerically by a gradient-based method, introducing a small relaxation parameter in order to assure differentiability of the penalty functional at zero. To be precise, instead of (10) we use

$$||\mathbf{v}||_{p,\varepsilon} := (\sum_{k \in I_0} (|v_k|^2 + \varepsilon)^{p/2})^{1/p} \tag{12}$$

combined with the hierarchical optimization strategy proposed in the same reference. The local search is implemented using the MATLAB routine *fmincon*, i.e. a sequential quadratic programming algorithm. In practice we also impose a priori lower and upper bounds for the unknown parameters in order to restrict the support of the posterior.

4.4. Stage 3: MCMC posterior exploration

From the first stage of the algorithm we have, based on the initial guess $\mathbf{x}_0$, identified the set I_0 of indices corresponding to unknown parameters that supposedly have small or no influence for reproducing the measured data. After elimination of a subset of these parameters via the sparsity promoting regularization in the second stage we denote the vector containing the remaining parameters $\mathbf{z} \in \mathbb{R}^s$, $s \leq m$. In the third stage we sample the posterior distribution of $\mathbf{Z}$. As third stage prior $\pi_{\text{pr},3}$ we choose a Gaussian distribution whose mean is the estimate obtained in the second stage and whose variance is chosen according to our (subjective) confidence in this estimate. Furthermore, we restrict the support of the posterior density by imposing the a priori lower and upper bounds for the remaining unknown parameters. For this reduced model the conditional mean estimate $\mathbf{z}_\text{CM}$ of $\mathbf{z}$ as well as confidence interval estimates for the parameter estimates are computed. That is, we give approximate answers to the questions *"Given the measured data and the third stage prior information, what is the most probable value of Z?"* and *"In what interval are the values of Z with P% probability, given the second stage prior and the measured data?"*, respectively.

As the reduced model is still high-dimensional, numerical integration has to be carried out using a Monte Carlo method. Given a sequence of independent draws $\mathbf{z}_1, \ldots, \mathbf{z}_K \in \mathbb{R}^s$, all distributed according to the third stage posterior $\pi_3(\mathbf{z}|\mathbf{y}_\text{m})$, it follows from the law of large numbers that

$$\frac{1}{K} \sum_{k=1}^{K} \mathbf{z}_k \to \mathbf{z}_\text{CM} \tag{13}$$

as $K \to \infty$. In practice, due to the fact that the posterior density is only known up to the normalizing constant in the denominator, it is impossible to generate independent samples. However, the above convergence remains valid if the samples are drawn from an ergodic chain having the posterior as its equilibrium distribution. Moreover since one can expect strongly correlated parameter combinations, it is worthwhile to consider some highly efficient adaptive MCMC sampler rather than standard Gibbs or Metropolis Hastings. In this study we use the delayed rejection adaptive (DRAM) variant of the latter, introduced by Haario et al. [15].

5. NUMERICAL EXPERIMENT

The biological setting to which our numerical experiment corresponds, spans the time period of the *cleavage cycle* 13[1]; from the end of the 12th mitotic division ($t = 0.0$ min) to the beginning of the 13th ($t = 16.0$ min). For our numerical experiment we choose $N_\text{m} = 4$ and observation times $t_1 = 1,\ t_2 = 5,\ t_3 = 9,\ t_4 = 13$.

In order to test our numerical scheme, we compute synthetic measurement data using our forward solver with parameters $R_a = 0.18$, $m_a = 0.15$, $h_a = 0.1$, $\lambda_a = 0.08$, $D_a = 0.1$, $a = 1 \ldots, n$, $W_1^1 = W_1^5 = W_2^2 = W_3^1 = W_5^3 = W_6^6 = 1.0$ and $W_1^3 = W_2^1 = W_3^6 = W_4^4 = 0.5$. The remaining entries of $\mathbf{W}$ are set to 0, representing the regulation parameters with no interaction. As initial conditions for the concentrations we prescribe a sinosoidal function for genes 1, 3 and 5 and a cosinusoidal function for gene 6. Initial concentrations for genes 2 and 4 are

[1] Periods between two consequent mitotic divisions

parameter	initial	lower	upper
R_a	0.15	0	0.25
m_a	0.15	0	0.25
h_a	0.1	0	0.15
λ_a	0.05	0	0.1
D_a	0.1	0	0.15
W_a^b	0.5	0	$\begin{cases}1.25\\0.5\end{cases}$

Table 1. Initial value and a priori lower and upper bounds for $a = 1, ..., 6$, where for the regulation parameters the second row lists those indices such that W_a^b has no interaction.

assumed to be equal to zero. Finally Gaussian noise with mean zero and standard deviation 1% of the maximum norm of the computed measurement vector is added. The initial values and a priori lower and upper bounds for the unknown parameters are given in Table 1.

For the sparsity promoting regularization we choose $\alpha = 0.1$, $p = 0.5$ and $\varepsilon = 10^{-3}$. In the third stage of the algorithm, we compute $K = 20000$ samples after a burn-in phase of 1000 samples.

In our numerical experiment we have observed, that, given sufficiently good prior knowledge, the sparsity promoting MAP estimate combined with a priori sensitivity analysis yields a good way to eliminate non-identifiable parameters. The normalized sensitivities computed in the first stage are plotted in figure 1. As it turns out the parameters R_a show a high sensitivity with respect to the given measurement data. On the other hand, the sensitivity of the diffusion coefficients is below the user-defined threshold. Notice moreover that the sensitivity of several regulation parameters is also below the threshold. In fact the normalized sensitivity for all of the regulation parameters is below 0.5, indicating that the objective functional is only moderately sensitive to variations in these parameters.

Figure 2 shows the result of the optimization carried out in the second stage. Again a user-defined threshold is employed in order to decide which unknowns should be dropped from the model. In the example presented here the dimension and at the same time the ill-posedness of the problem could be reduced significantly as 33 parameters were eliminated. In particular the resulting reduced gap gene model does not contain any diffusion terms. Notice that such diffusion-less approximations have been proposed in the biological literature previously. However, when it comes to other parameters, such as regulation parameters, it requires some biological expertise to decide whether a reduced model is justifiable. If it turns out it is not, the a priori bounds should be adjusted accordingly.

Figure 3 shows the conditional mean estimate for the posterior of the reduced system. Confidence intervals and the whiskers yield some additional information about the reliability of the estimate. As one would expect from the first stage sensitivity analysis the parameters R_α in the reduced model can be inferred from the measured data in quite a stable manner, while the uncertainty for the estimates of the regulation parameters is significantly larger. Although we can not go into detail here due to space restrictions, it should be emphasized that the chains produced via MCMC sampling include plenty of additional information such as cross-correlations, etc..

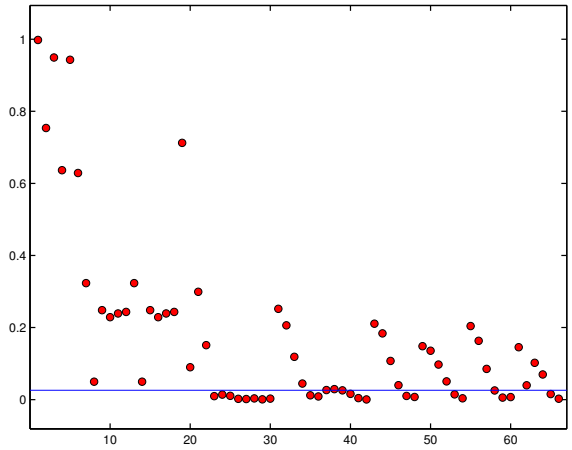


Figure 1. First stage: Sensitivities of the unknown parameters with respect to the given measurement data at the initial guess $\mathbf{x}_0$. The horizontal line shows the user-defined threshold.

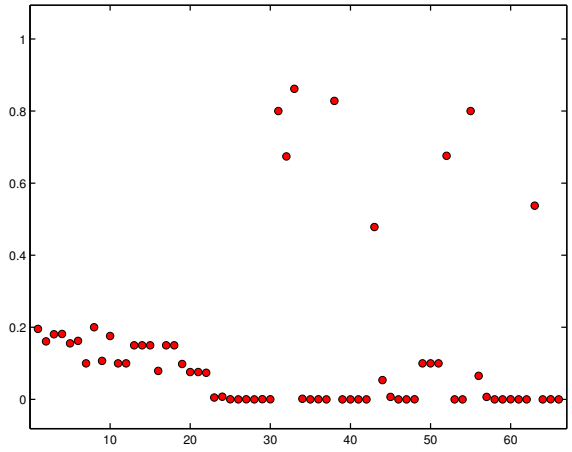


Figure 2. Second stage: MAP estimate for the full set of unknowns with the initial guess and lower and upper bounds from table 1.

Finally figure 4 depicts the posterior plot for the concentration of gene 3 showing that the reduced model can indeed produce accurate predictions.

6. CONCLUSIONS

Within the framework of Bayesian inversion it is possible, given sufficiently good prior knowledge, to eliminate non-identifiable parameters from the full gap gene circuit model and to obtain parameter estimates for the resulting reduced model in a stable manner. It should be emphasized that all available prior information can be included into the proposed scheme in a rather explicit way. Adap-

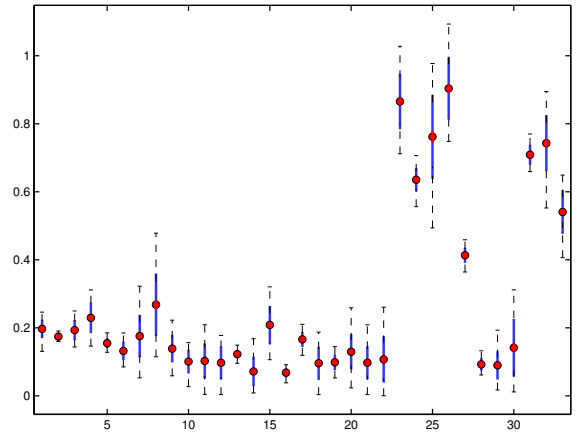


Figure 3. Third stage: Posterior mean for the reduced system with 95% confidence intervals. The whiskers extend to the most extreme datapoints.

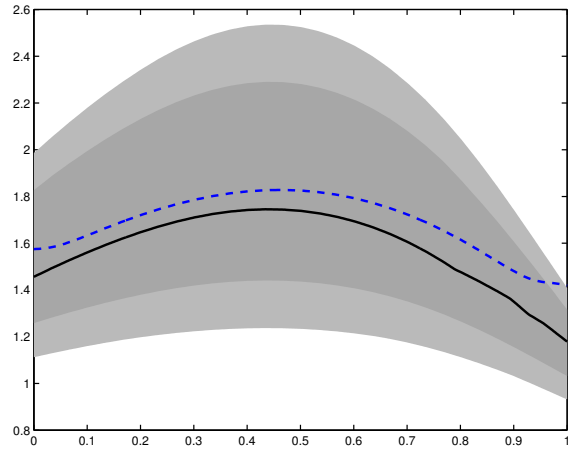


Figure 4. The solid line gives the posterior mean of the concentration of gene 3 at time $t = 16.0$. The dashed line corresponds to the 'true' concentration. Gray areas correspond to 90% and 95% posterior uncertainty, respectively.

tive MCMC sampling of the posterior distribution of the reduced model is an efficient way to assess the uncertainty inherent in experimental measurements as well as in the mathematical models and their computational implementation. For synthetic data it was shown that the reduced gap gene model can produce reasonable predictions for the experimentally measured concentrations.

We hope that the proposed scheme can be a valuable addition to established techniques, thus possible future work arising from this study would be a test with real data, using prior knowledge produced e.g. by the approach in [3]. In a similar directions we consider more general problems and models including several cleavage cycles, and delay parameters for the mitosis e.g. as described in [16, 17].

References

[1] H. Engl, M. Hanke, and A. Neubauer, *Regularization of inverse problems*. Kluwer Academic Publishers Group, Dordrecht, 1996.

[2] J. Kaipio and E. Somersalo, *Statistical and computational inverse problems*. Springer-Verlag, New York, 2005.

[3] M. Ashyraliyev, J. Jaeger, and J. Blom, "Parameter estimation and determinability analysis applied to drosophila gap gene circuits," *BMC Systems Biology*, vol. 2:83, 2008.

[4] P. Kuegler, E. Gaubitzer, and S. Mueller, "Parameter identification for chemical reaction systems using sparsity enforcing regularization - a case study for the chlorite - iodide reaction," *Journal of Physical Chemistry A*, vol. 12, pp. 2775–2785, 2009.

[5] H. Engl, C. Flamm, P. Kuegler, J. Lu, S. Mueller, and P. Schuster, "Inverse problems in systems biology," *Inverse Problems*, vol. 25, no. 12, 2009.

[6] P. W. Ingham, "The molecular genetics of embryonic pattern formation in *Drosophila*," *Nature*, vol. 335, pp. 25–34, Sep 1988.

[7] M. Akam, "The molecular basis for metameric pattern in the *Drosophila* embryo," *Development*, vol. 101, pp. 1–22, Sep 1987.

[8] C. Nusslein-Volhard and E. Wieschaus, "Mutations affecting segment number and polarity in *Drosophila*," *Nature*, vol. 287, pp. 795–801, Oct 1980.

[9] V. Foe and B. Alberts, "Studies of nuclear and cytoplasmic behaviour during the five mitotic cycles that precede gastrulation in drosophila embryogenesis," *The Journal of Cell Science*, vol. 61, pp. 31–70, 1983.

[10] J. Reinitz and D. Sharp, "Mechanism of eve stripe formation," *Mech. Dev.*, vol. 49, pp. 133–158, 1995.

[11] J. Jaeger, M. Blagov, K. Kozlov, E. Manu Myasnikova, S. Surkova, C. Vanario-Alonso, M. Samsonova, D. Sharp, and R. J., "Dynamic analyses of regulatory interactions in the gap gene system of drosophila melanogaster," *Genetics*, vol. 167, pp. 1721–1737, 2004.

[12] J. Jaeger, S. Surkova, M. Blagov, H. Janssens, D. Kossman, K. Kozlov, E. Manu Myasnikova, C. Vanario-Alonso, M. Samsonova, D. Sharp, and R. J., "Dynamic control of positional information in the early drosophila embryo," *Nature*, vol. 430, pp. 368–371, 2004.

[13] J. Lu, S. Muller, R. Machne, and F. C., "Smbl ode solver library: Extensions for inverse problems," *Proceedings of the fifth workshop of on computational systems Biology*, 2008.

[14] M. Grasmair, "Well-posedness and convergence rates for sparse regularization with sublinear l^q penalty term," *Inverse Probl. Imaging*, vol. 3, no. 3, pp. 383–387, 2009.

[15] H. Haario, M. Laine, and A. Mira, "Dram: Efficient adaptive mcmc," *Statistics and Computing*, vol. 16, no. 4, pp. 339–354, 2006.

[16] J. Jaeger, "The gap gene network," *Cell Mol. Life Sci.*, vol. 68, pp. 243–274, 2011.

[17] K. Becker, "A quantitative study of translational regulation in drosophila segment determination," Master's thesis, Institute of Genetics, University of Mainz, 2013.

Acknowledgements
Both authors wish to thank the *"Center for Computational Sciences in Mainz"*, and N.S. the *"Alexander von Humbold Foundation"* for their support during the development of this work.
Both authors express their thanks to Kolja Becker for his invaluable assistance concerning the biological background of this work.

PERTURBATION PROPAGATION IN BOOLEAN NETWORKS WITH LOCAL STRUCTURES

Tero Soininen[1], *Matti Nykter*[1,2] *and Juha Kesseli*[1]

[1]Department of Signal Processing, Tampere University of Technology,
P.O. Box 553, FI-33101 Tampere, Finland
[2]Institute of Biomedical Technology,
FI-33014 University of Tampere, Finland
tero.soininen@tut.fi

ABSTRACT

In this work, we present a probabilistic model of partially annealed network dynamics in the case of Boolean networks with numerous identical feed-forward loop motifs. The results obtained with our model are compared to iterated Derrida maps and data from simulations done on partially annealed Boolean networks. This comparison shows that in most cases our solution surpasses the accuracy of the Derrida map when estimating perturbation propagation. However, our model is based on the simplification that the bias of the network is not changed on average by the perturbations. Hence, there are cases in which more complex models are needed.

1. INTRODUCTION

Complex networks can be found practically everywhere around us: in cells, social interactions, climate, electricity distribution and the Internet. The study of complex networks, being a very widely applicable field of research, has gained a lot of interest in the last couple of decades. This has given rise to many approaches of modeling these networks, e.g., differential equations, Boolean networks and Petri nets [1].

Boolean networks can be useful when studying network level dynamical properties in large networks. Boolean network is a set consisting of nodes $1...N$, the states of the nodes $\boldsymbol{x} \in \mathbb{B}^N$ and the update rules for the nodes $\boldsymbol{f} : \mathbb{B}^N \to \mathbb{B}^N$. The states are updated at discrete time steps according to the update rules, that is, $\boldsymbol{x}(t+1) = \boldsymbol{f}(\boldsymbol{x}(t))$. The functions can be presented as a vector, e.g., $[0001]$, where the vector is the truth table column of the function in ascending binary order. The variables $x_i(t)$ that affect the state of $x_j(t+1)$ are called the inputs of node j and the number of these inputs is the in-degree of node j. As one would assume, this simple structure makes Boolean networks rather simple from the computational perspective as well.

In the field of computational systems biology the idea of using Boolean networks has a long tradition. In 1969 Stuart Kauffman proposed the use of Random Boolean networks (RBNs) for modeling gene regulatory networks. In these so-called Kauffman networks the update functions of the nodes and the connections between the nodes are chosen at random, yet the in-degree of the nodes is kept constant [2]. These networks consisting of N nodes and having a constant in-degree of K for all of the nodes are also often referred to as NK networks or Kauffman networks [3]. Indeed, Kauffman's article showed the value of Boolean networks for theoretical biologists; a theoretical model, such as a Boolean network, can give insight into the principles of dynamical behaviour of a network even if the detailed network structures remain unclear.

Different properties of Boolean networks can be connected with observables from real networks [2]. One of these properties is the bias, $b(t)$, of the network, i.e., the proportion of ones in the network state and especially the steady state value b^*. In the models of this article we assume that the networks have reached the bias steady state. For RBNs the bias steady state corresponds to the proportion of ones in the functions of the nodes. This function bias is denoted by b_f.

Propagation of perturbations is particularly interesting since it can be considered to characterize the response of the system to internal perturbations and outside inputs. Individual networks, however, have complicated nonlinear dynamics and an exhaustive analysis of their state space is typically impossible. To overcome these problems there are tools such as the annealed approximation, which predict the dynamics using a probabilistic approach which assumes that the effect of local structure on the dynamics can be neglected [4].

Unfortunately, real regulatory networks are not random; studies indicate that some local structures, called motifs, appear more often than what would be expected in a random network [5]. These local structures have an effect on the dynamical properties of the network that they are a part of. To study these effects a partially annealed approximation has been proposed [6]. Partially annealed approximation means that the connections that are not part of a motif are shuffled at each time step. As feed-forward loops introduce a kind of a memory in the network dynamics, the standard annealed analysis will not, in general, be sufficient. In this article, the aim is to find an ana-

lytical model for propagation of perturbations in networks with considerable amounts of local structures, particularly feed-forward loops.

2. LOCAL STRUCTURES AND BIAS IN BOOLEAN NETWORKS

This section introduces two models for approximating the bias steady state distribution of a C1-FFL motif from [6] and then replicates the results obtained in the aforementioned article. The analysis of bias fixed points is needed as a first step towards perturbation calculations.

In this article we use the same naming convention as in Alon's article [5] for feed-forward loops. Our main example, C1-FFL is a feed-forward loop, which can be modeled in Boolean networks using 3 nodes. The circles represent nodes and the arrows are connections between the nodes. The inputs for the node A are called input nodes for the feed-forward loop and they can come from whichever nodes in the network. Node B simply replicates the state of A at the previous time step and node C expresses the $[0001]$ function of nodes A and B.

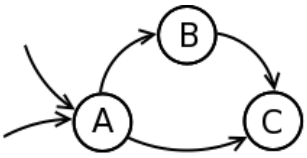


Figure 1. Illustration of a Boolean network representing a C1-FFL motif.

2.1. Markov chain models

The state of the motif at $t+1$ is defined by a transition matrix $\boldsymbol{M}$ and the state of the motif at the previous time step t. The transition matrix contains the probabilities for the state transitions and it can be deduced from a state transition diagram. In the transition matrix the row of the element depicts the current state and the column the state at the next time step. So, for example, element $m_{2,1}$ contains the probability of transitioning from state 001 to 000, where the bits indicate the state of nodes A, B and C, respectively.

When $\boldsymbol{p}$ is a vector that contains probabilities for the states of the motif in an ascending binary order, i.e., the first element is the probability that the motif is in state 000, the second element for state 001 and the last for 111, the update rule can be expressed as

$$\boldsymbol{p}^T(t+1) = \boldsymbol{p}^T(t)\boldsymbol{M}, \tag{1}$$

where

$$\boldsymbol{M} = \begin{pmatrix} u_0 & 0 & 0 & 0 & \bar{u_0} & 0 & 0 & 0 \\ u_1 & 0 & 0 & 0 & \bar{u_1} & 0 & 0 & 0 \\ u_1 & 0 & 0 & 0 & \bar{u_1} & 0 & 0 & 0 \\ u_2 & 0 & 0 & 0 & \bar{u_2} & 0 & 0 & 0 \\ 0 & 0 & u_1 & 0 & 0 & 0 & \bar{u_1} & 0 \\ 0 & 0 & u_2 & 0 & 0 & 0 & \bar{u_2} & 0 \\ 0 & 0 & 0 & u_2 & 0 & 0 & 0 & \bar{u_2} \\ 0 & 0 & 0 & u_3 & 0 & 0 & 0 & \bar{u_3} \end{pmatrix}. \tag{2}$$

In the transition matrix M, u_i denotes the probability that the node A of the motif is in state 0 at the next time step when i is the number of nodes in state 1 in the motif. And for the case that the aforementioned node is in state 1, we denote $\bar{u_i} = 1 - u_i$. For the first model u_i is considered to be constant, i.e., $u_i = b_f, \forall i$. To calculate u_i for the second model we need to know the proportion of nodes belonging to the motifs α and the probability $b_{f_{bg}}$ that a background node is in state 1. Now,

$$u_i = ((1-\alpha)(1-b_{f_{bg}}) + \alpha\frac{3-i}{3})^2. \tag{3}$$

It is shown in [6] that both of the models in this Section do indeed converge to a steady state $\boldsymbol{p}^*$ and that this steady state can be calculated in a rather simple manner from the eigenvector $\boldsymbol{v}$ associated to eigenvalue $\lambda = 1$ of the transition matrix $\boldsymbol{M}$,

$$\boldsymbol{p}^* = \frac{\boldsymbol{v}}{\|\boldsymbol{v}\|}. \tag{4}$$

2.2. Results

The results shown here are from simulations run on Matlab and they consist of 100 Boolean networks without annealing and another 100 networks with annealing. Each network was built out of NK networks of 2400 nodes with $K = 2$ by rewiring and changing the functions of 1200 nodes so that 400 C1-FFL motifs with $f_A = [0111]$ were formed. These networks were run for 10000 time steps to reach a bias steady state. Then the mean of the states of the motifs was calculated. The proportion of nodes belonging to the motifs was $\alpha = 0.5$. The function bias of the background nodes was $b_{f_{bg}} = 0.7$. The theoretical results are calculated from the eigenvectors of the transition matrices as shown in Equation 4.

The results are plotted in a bar graph in Figure 2 so that it shows the steady state distribution of the states of the motif. In the graph the height of the column is the probability that a motif is in that state. The results from the simulations done for this article repeat the ones obtained in [6].

The first model does not quite succeed in estimating the steady state distribution, but when looking at the results of the second model, we can see that they are significantly closer to those of the annealed networks. Hence, the second model can be used to estimate the results from such networks. When the annealing is dropped the results

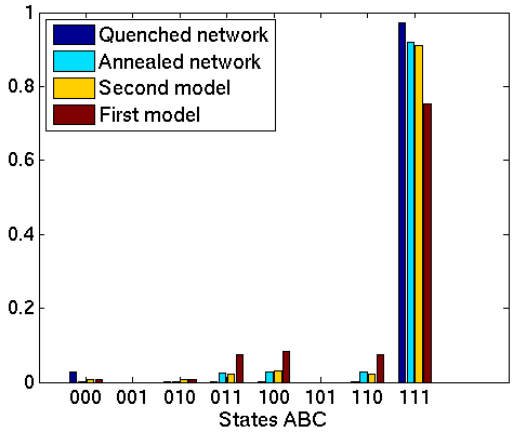


Figure 2. The above bar plot shows the steady state distributions from two simulations and two different analytical models. The simulation results are averaged from 100 networks of 2400 nodes containing 400 C1-FFL motifs for which $f_A = [0111]$. Two simulations were done: without and with annealing. State ijk means that node A is in state i, node B in j and node C in k.

from the networks are different. This difference should get smaller as the size of the network goes up, since the statistical properties of the quenched network get closer to those of the annealed one. A more thorough analysis of these results is found in [6].

3. LOCAL STRUCTURES AND PERTURBATION PROPAGATION IN BOOLEAN NETWORKS

This section first introduces the so-called Derrida map and illustrates the need for another perturbation propagation model to be used in the case that the network includes a considerable amount of motifs. Then a model for networks with feed-forward loops is introduced and the results obtained with that model are compared to the ones obtained with the Derrida maps.

3.1. The Derrida map

The usual way of illustrating the propagation of perturbations in a Boolean network is a Derrida map. The Derrida map shows the average Hamming distance $\rho(t+1)$ of two identical networks' states at time $t+1$ as a function of the distance $\rho(t)$ at time t. It is named after B. Derrida who introduced it to study the evolution of overlaps between configurations in RBNs [7].

Let f^n denote a function composition, that is, function f is iterated n times. Now, say $f : [0,1] \rightarrow [0,1]$ defines a Derrida map: $\rho(t+1) = f(\rho(t))$. For RBNs, when we measure the average propagation of perturbations, Equation 5 holds.

$$\rho(t+n) = f^n(\rho(t)) \tag{5}$$

But as can be seen in the example given in Figure 3, when we add local structures to the network, the iterated mappings do not match those acquired from the simulations, and thus, the Derrida map can not be used to predict the size of the perturbation avalanches over multiple time steps. That is, Equation 5 does not apply in this case.

The iterated maps in Figure 3 are formed by iterating a second degree polynomial that was fitted to the data from a simulated network. For the Derrida map to work, the iterated maps at $t+k, k > 1$ should match those from the simulations. In fact, the behaviour of the iterated maps in this case is the opposite of what is actually observed. The iterated maps rise higher, while the simulations show a dip below the diagonal after a few time steps. This is a radical difference, as the iterated maps suggest that the perturbation grows bigger with time on average, when in reality it would seem to die out eventually, which agrees with the result in [6] that C1-FFL increases the stability of the network.

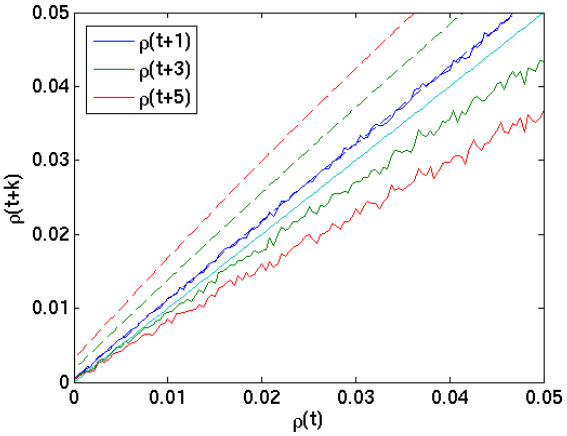


Figure 3. Derrida map of networks with C1-FFL motifs. Iterated maps are shown with the dashed line and the simulated data with solid line. The results above are averaged from 100 partially annealed networks with 2400 nodes and 400 C1-FFLs with $f_A = [0111]$. A second degree polynomial was fitted to the perturbation propagation data after one time step and this polynomial was then iterated. As seen above, the iterations do not match the results from the simulations.

3.2. Perturbation model for feed-forward loops

To address the shortcomings of the previous Derrida map approach, we suggest a partially annealed model to approximate the local structure effects that the feed-forward loops introduce in the dynamics. The model divides the nodes of the network into four parts: the A, B and C nodes in the motifs and the background nodes denoted by bg. Let us assume that we have a network with a considerable amount of C1-FFL motifs and the initial perturbation happens at time t. Now $\rho_A(t) = \rho_B(t) = \rho_C(t) = \rho_{bg}(t) = \rho(t)$ and the updating of these partial perturbations happens according to the Equations 6–9. In order to calculate the partial perturbations we also need to know

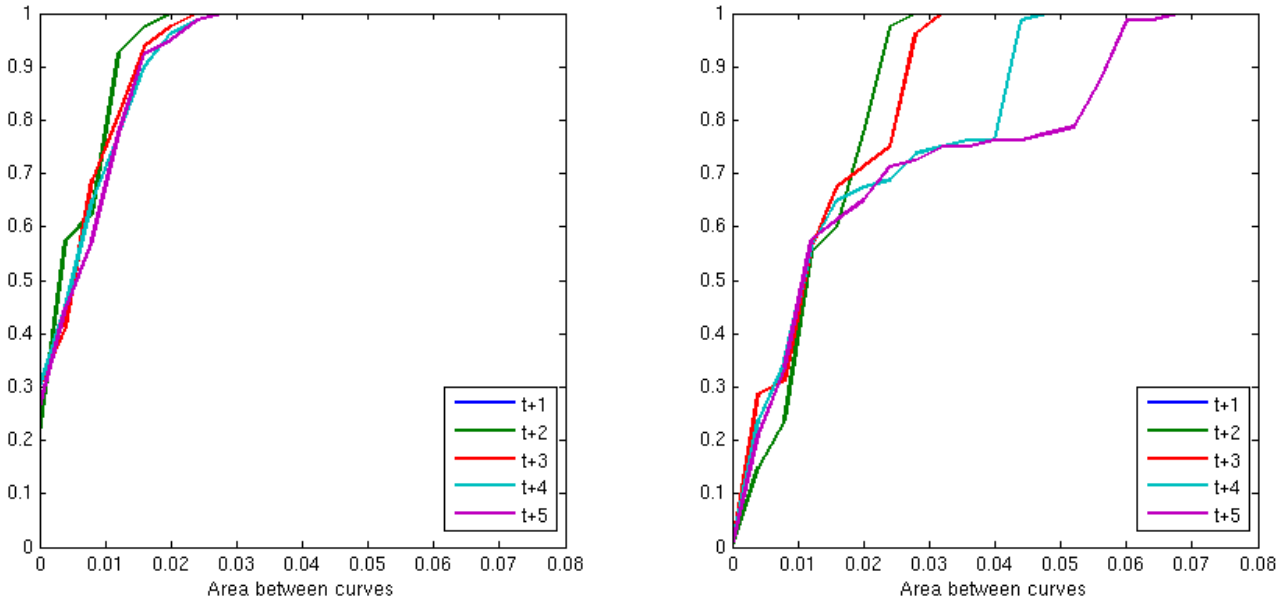


Figure 4. Cumulative error distribution at different time steps, when the error measure is the area between the mappings of the model estimate and the simulated results. Results of the Derrida map on the right and the model of Section 3.2 on the left.

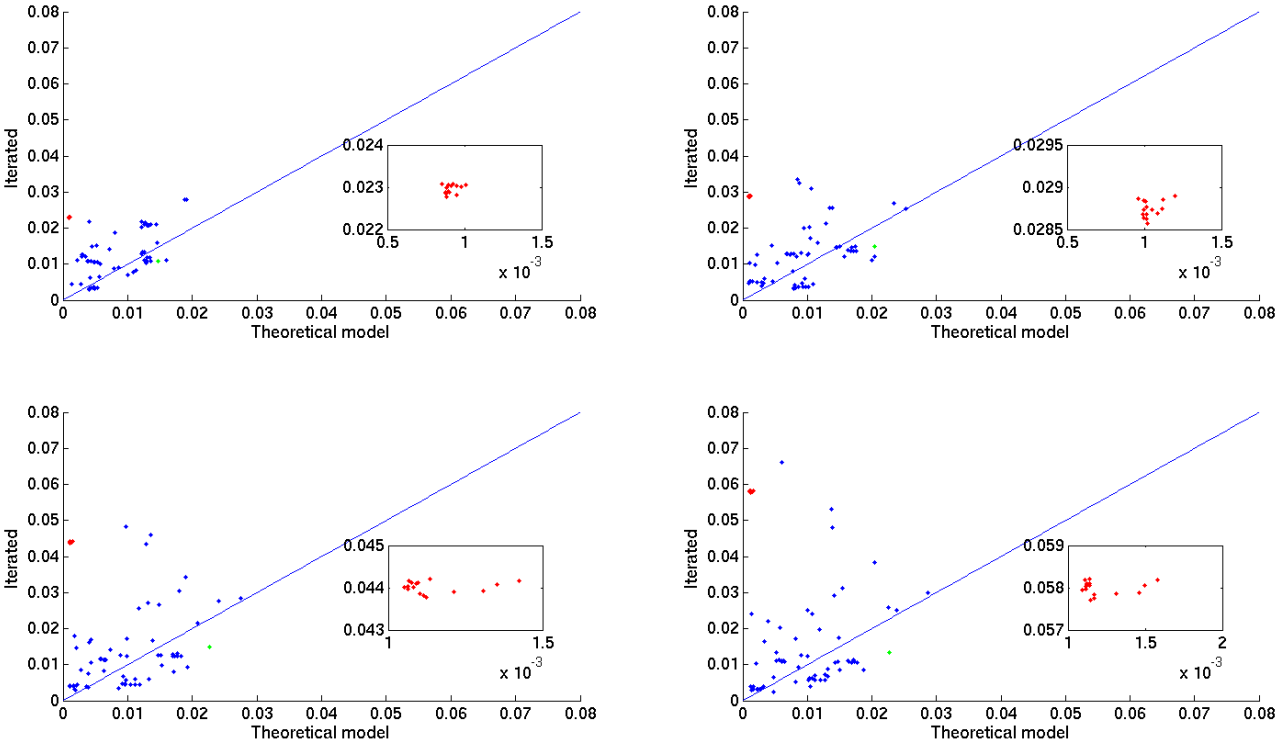


Figure 5. Scatter plot of the errors when the error measure is the area between the mappings. Time advances from left to right, top to bottom, beginning at $t + 2$ in the top left and ending at $t + 5$ in the bottom right. The inset is a zoom in on the cluster of red markers. This cluster is comprised of those networks for which $f_A = [0110]$ or $[1001]$. The instance marked in green is seen in Figure 6.

the bias steady states of the nodes belonging to the motif b^*_A, b^*_B, b^*_C and the bias steady state of the background nodes b^*_{bg}. These biases can be obtained from simulations or for example with the models of Section 2.1. Then, knowing the amount, α, of nodes belonging to the motifs we can calculate the total size of the perturbation given by Equation 10.

$$\rho_A(t+1) = 2(1-b^*)\rho(t)(1-\rho(t)) + \\ + (b^{*2} + (1-b^*)^2)\rho(t)^2 \quad (6)$$

$$\rho_B(t+1) = \rho_A(t) \quad (7)$$

$$\rho_C(t+1) = b^*_A(1-\rho_A(t))\rho_B(t) + b^*_B\rho_A(t) \\ (1-\rho_B(t))(b^*_A b^*_B + (1-b^*_A)(1-b^*_B)) \\ \rho_A(t)\rho_B(t) \quad (8)$$

$$\rho_{bg}(t+1) = (2\rho(t)(1-\rho(t)) + \rho(t)^2)2b^*_{bg}(1-b^*_{bg}) \quad (9)$$

$$\rho(t+1) = \frac{\alpha}{3}(\rho_A(t+1) + \rho_B(t+1) + \rho_C(t+1)) + \\ + (1-\alpha)\rho_{bg}(t+1) \quad (10)$$

The equations for approximating the propagation of perturbations are obtained by looking at how the functions in the feed-forward loops handle perturbations in their inputs. For example, when we have a C1-FFL with $f_A = [0111]$, one perturbation in the input nodes advances to node A, if and only if, both of the inputs are in state 0. If both of the input nodes are flipped, then the perturbation advances only if the input nodes are both in state 0 or both in state 1. Perturbation in node A advances always to node B and the possibilities for perturbations advancing to node C can be deduced from the truth tables.

3.3. Results

The results in this section are obtained using the two models introduced in Sections 3.1 and 3.2. Equations similar to 6–9 were written for each of the 80 possible motifs, with biases from numerical simulations(not shown due to space constraints). These results are then compared to those from simulations from a total of 8000 different networks. This is because 80 different feed-forward loops were considered and the simulated results are always averages from 100 different networks. Each network was created from the so-called NK-networks with $N = 2400$ and $K = 2$ by adding 400 motifs. The networks were then updated to reach the bias steady state, a copy was made and a perturbation was introduced. Then, the average Hamming distance of these two networks was calculated at five time steps after the initial perturbation.

Two different measures were used in the comparison of the Derrida map and the other model, the first one being the area between the curves of the simulated data and the estimate; in this measure, a smaller value means a better result. The second measure(not shown), which was the maximum difference between the estimate and the simulated model, gave similar results as the area measure.

A scatter plot of the area errors with all the different feed-forward loops was also drawn. In the scatter plot, a clear cluster can be seen formed by those feed-forward loops for which $f_A = [0110]$ or $[1001]$. This cluster was coloured red to highlight the points shown in the zoomed inset. Also, one of the instances for which the Derrida map gives a smaller error than the other model is observed more thoroughly in Figure 6. This instance marked with green in the scatter plot is a C3-FFL for which $f_A = [0111]$, $f_B = [01]$ and $f_C = [1000]$.

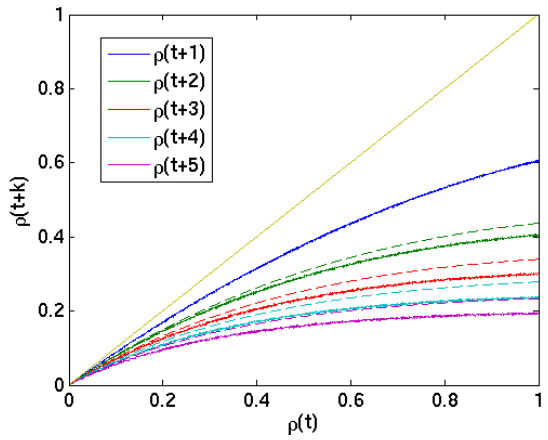


Figure 6. Propagation of perturbations according to the probabilistic model and simulated results averaged from 100 networks of 2400 nodes containing 400 C3-FFLs with $f_A = [0111]$. The model results are shown as the dashed line and the simulated results as the solid line.

4. CONCLUSION AND DISCUSSION

This work started with the replication of the results in [6] for the bias steady state in Boolean networks with local structures. Then, a model for estimating the propagation of perturbations in Boolean networks with local structures was introduced. This model, based on applying partial annealing, could be seen to improve on standard Derrida map analysis in most cases where different feed-forward loops were present in abundance.

Looking at the error distributions in Figure 4 it appears quite clear that on average the probabilistic model predicts the propagation of perturbations in Boolean networks with local structures better than the Derrida map, and the scatter plot in Figure 5 supports this claim. Although not the case for all the feed-forward loops, most of them favour the presented probabilistic model.

One of the most interesting phenomena is the cluster that the feed-forward loops with $f_A = [0110]$ or $[1001]$ create in the scatter plot. This cluster is located rather far above the diagonal, i.e., the Derrida map fails completely at estimating the propagation of perturbations in these cases, whereas the probabilistic model is very close to the simulation data.

However, there were also points below the diagonal. These points illustrate those FFLs where the Derrida map

performed better than the probabilistic map. One of these was the C3-FFL with $f_A = [0111]$ for which the annealed approximation and the simulated results are shown in Figure 6. The most likely cause for these shortcomings is the average change in network bias that happens due to the introduction of large random perturbations in the network. Overcoming this problem would, however, require a more complex model for the approximation than what was described in this article.

5. ACKNOWLEDGMENTS

We acknowledge the funding from Emil Aaltonen Foundation and the Academy of Finland project 251937.

6. REFERENCES

[1] S. H. Strogatz, "Exploring complex networks," *Nature*, vol. 410, pp. 268–276, Mar. 2001.

[2] S. A. Kauffman, "Metabolic stability and epigenesis in randomly constructed genetic nets," *Journal of Theoretical Biology*, vol. 22, pp. 437–467, Mar. 1969.

[3] E. Dubrova, M. Teslenko, and A. Martinelli, "Kauffman networks: analysis and applications," in *Computer-Aided Design, 2005. ICCAD-2005. IEEE/ACM International Conference on*, Nov. 2005, pp. 479 – 484.

[4] B. Derrida and Y. Pomeau, "Random networks of automata: A simple annealed approximation," *Europhysics Letters*, vol. 1, pp. 45–49, Jan. 1986.

[5] U. Alon, "Network motifs: theory and experimental approaches," *Nature Reviews Genetics*, vol. 8, pp. 450–461, Jun. 2007.

[6] V. Picard, "Modeling of topological effects in biological networks," Internship report 2010. Online; Accessed 2013, Jan 30, http://www.irisa.fr/dyliss/system/files/public/vpicard/m1_report.pdf.

[7] B. Derrida and G. Weisbuch, "Evolution of overlaps between configurations in random Boolean networks," *Journal de Physique*, vol. 47, pp. 1297–1303, Aug. 1986.

ANALYSIS OF FACTORS AFFECTING THE GROWTH OF OIL BODIES IN *A. THALIANA* SEEDS : USE OF ORDINARY LEAST SQUARES AND QUANTILE REGRESSION

Ghassen Trigui[1,2], *Martine Miquel*[2], *Bertrand Dubreucq*[2], *Olivier David*[1], *Alain Trubuil*[1,*]

[1]Institut National de la Recherche Agronomique, UR0341 MIA
F-78352 Jouy en Josas, France
[2]Institut National de la Recherche Agronomique, UMR1318 INRA-AgroParisTech,
Route de Saint-Cyr, 78026 Versailles, France
Alain.Trubuil@jouy.inra.fr (*Corresponding author)

ABSTRACT

The sub-cellular organelles called oil bodies (OBs) are lipid-filled quasi spherical droplets produced from the endoplasmic reticulum (ER) and then released into the cytoplasm during seed development. It is believed that an OB grows by coalescence with other OBs and that its stability depends on the composition in oleosins, major proteins inserted in the hemi membrane that covers OBs. Five oleosin proteins, namely S1 to S5, were discovered. The size of OBs evolves during the first steps of seed development, but is also contingent upon their protein complement. Individualized volumes of OBs were extracted from confocal microscopy images of embryos from different genotypes of *A. thaliana* seeds at different days after flowering (DAF). Models based on ordinary least squares (OLS) and quantile regression (QR) estimators were proposed to analyze the factors associated with the growth of OBs. Whatever the estimator, S1 oleosin showed a significant effect in reducing the volume of OBs while S4 contributed to its increase. S3 was shown to act by reducing OB volume (p-values < 0.001, OLS) but only in higher ranges of volume using QR. Over all selected quantiles (in QR) and within OLS, a significant synergistic interaction between S3 and S4 was shown, while a null interaction between S1 and S4 was clearly shown within the QR in low and high (p-value = 0.69, 0.1-quantile and 0.57, 0.75-quantile respectively) volumes of OBs, as well as in OLS (p-value $= 0.99$).

1. INTRODUCTION

In most eukaryotic organisms, storage lipids are deposited in stable sub-cellular structures named lipid or oil bodies (OBs). These structures are produced from the endoplasmic reticulum (ER) and then released into the cytoplasm during seed development [1]. OBs differ from one species to another and between kingdoms, particularly by their composition in neutral lipids and protein complements [2]. OBs are heterogeneous in size and in number, and exhibit growth dynamics different from one organism to another and from one cell type to another. Indeed, according to its energy needs, the cell adopts the configuration size and/or number of lipid bodies which optimizes storage capacity, production, and consumption. However, the growth mechanisms of lipid bodies are still unclear, even if different hypotheses have been proposed [3]. In seeds, a family of proteins called oleosins have been identified on the surface of OBs[4]. These proteins seem to play an important role in the dynamics of OBs, and have been suggested to act as "stabilizers" preventing OBs coalescence. Studies of OBs in *A. thaliana* embryos deficient in oleosins showed impaired lipid and protein accumulation accompanied with a delay in germination, and abnormal over-sized OBs [5]. In statistical analysis, the conditional mean of observations is often estimated by ordinary least squares (OLS) in an analysis of variance. This may not be informative enough, particularly in the case of skewed distributions, where the few values on the tails of the distribution are neglected. These values, despite their small number, represent important information when considering for example the volume of individuals in a population. In this situation, analysis of quantiles by quantile regression (QR) can be an alternative solution to extract information from these values [6]. The advantage of this method is that we can track the range of data in which the effect of associated covariates is impacting. Originally developed for econometrics [7], quantile regression is more and more used in biostatistics and life science fields [8]. The aim of this paper is to study the factors involved in the growth of OBs through ordinary least squares (OLS) and quantile regression (QR).

2. METHODS

2.1. Data acquisition

Arabidopsis thaliana wild type and oleosin mutant plants defective for one or several oleosins were grown from seeds on soil in a greenhouse. Upon flowering, flowers and subsequent developing siliques were tagged daily until 12 days after flower opening. Siliques for each stage of development were sampled and dissected to remove developing seeds. Seeds were spread on a glass slide, incubated with Nile Red, a neutral lipid stain at a final concentration of 1μg/ml in a 60% glycerol solution. Embryos were removed from the seed teguments by gently pressing

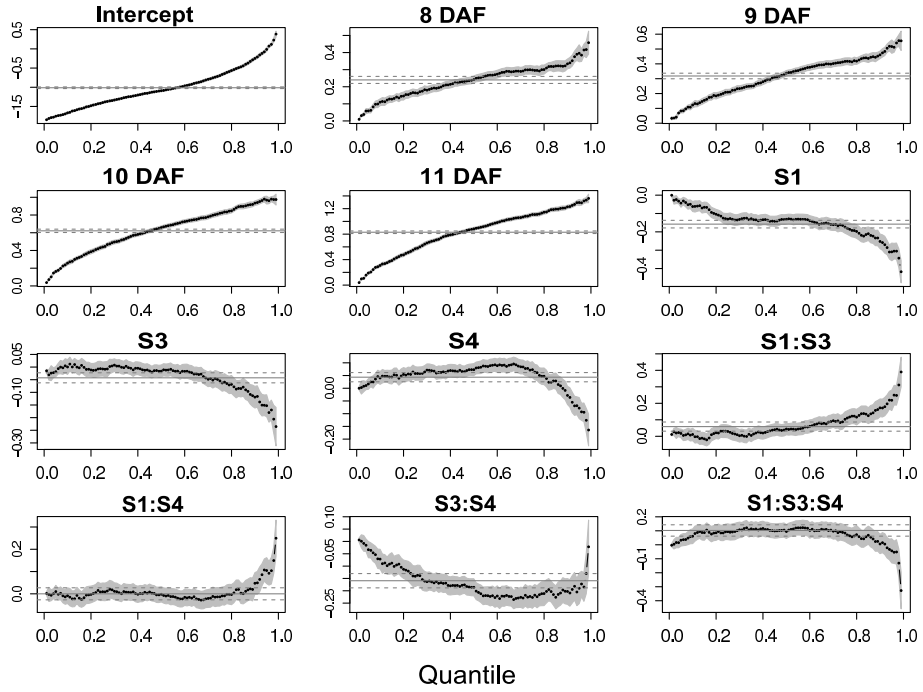


Figure 1. OLS and QR coefficients estimation : Coefficients estimation showing the effect of day, oleosin, and interactions between oleosins. OLS estimator coefficients (horizontal solid lines) and QR estimator coefficients (dashed dotted lines) are presented with their 95% confidence interval.

seeds between slide and coverslip and observed after 30 min of incubation in the dark. *Arabidopsis* oleosin null mutants are available for 3 oleosins, *S1* (At3g01570), *S3* (At4g25140) and *S4* (At5g40420) (ref. NASC). Double mutants (*s1s3*, *s1s4*, *s3s4*) and a triple mutant (*s1s3s4*) have been generated in the laboratory.

3D images of dissected *Arabidopsis* embryos were acquired using a LEICA SP2(AOBS) confocal microscope with a spatial resolution of $(0.09\,\mu m \times 0.09\,\mu m \times 0.16\,\mu m)$ in the (x, y, z) referential. Third dimension was obtained by scanning the sample through the z axis, providing a sequence of 2D images corresponding to the fluorescence emitted from the focal plane. Images were first filtered using ND-SAFIR, a software for denoising n-dimensional images especially dedicated to microscopy image sequence analysis [9], then segmented through a pipeline we designed using several algorithms from Avizo-Fire (Burlington, USA), a 3D image processing software.

2.2. Statistical analysis

A total of 112 three-dimensional images from independent samples were analyzed. Each image corresponds to one of the 8 genotypes, observed at one of the 5 development stages namely day 7, 8, 9, 10, and 11 days after flowering (DAF). At least two or three samples for each couple (genotype-day) were used, and from which individualized volumes of OBs were extracted. Volumes were classified on subsets of (genotype-day). In total, 50,379 OB volumes were quantified.

In the model we developed, we made the assumption that the volume of OB was affected both by the three oleosin factors (S1, S3 and S4), and the day factor. Each oleosin factor was labeled by an index, noted i, j, and k for S1, S3, and S4, respectively. Each index had two levels corresponding to the presence or the absence of the oleosin. The day factor contained five levels, noted by the label t, corresponding to 7, 8, 9, 10, and 11 DAF. Volumes V of OBs were transformed to their decimal logarithm $Log_{10}(V)$ in order to verify normality assumption, and equality of variance conditions needed for OLS. The model is expressed as:

$$\begin{aligned} y^n_{ijk,t} \quad = \quad & \textbf{intercept} + \textbf{Day}_t + \textbf{S1}_i + \textbf{S3}_j + \textbf{S4}_k \\ & + \textbf{S1:S3}_{ij} + \textbf{S1:S4}_{ik} + \textbf{S3:S4}_{jk} \\ & + \textbf{S1:S3:S4}_{ijk} + \varepsilon^n_{ijk,t} \end{aligned}$$

Where: $y^n_{ijk,t}$ is the value of $Log_{10}(V)$ of the n_{th} OB on the population with oleosin labels (ijk) at day level t. $\textbf{S1}_i$, $\textbf{S3}_j$ and $\textbf{S4}_k$ are respectively the main effects of factors S1, S3, and S4. $\textbf{S1:S3}_{ij}$, $\textbf{S1:S4}_{ik}$, and $\textbf{S3:S4}_{jk}$ are the effects of double interactions between oleosins. $\textbf{S1:S3:S4}_{ijk}$ is the effect of the triple interaction between S1, S3 and S4. Last, $\varepsilon^n_{ijk,t}$ is the error term which is supposed to follow a normal distribution with mean 0 and variance σ^2. The OLS estimator uses the minimization of the sum of squares to fit predictions to observations:

$$\widetilde{\theta} = argmin_\theta \left[\sum_{n=1}^{N} [y^n - f(\theta, n)]^2 \right] \qquad (1)$$

Where y^n denotes the responses for observation n , θ the vector of parameters to be estimated, and $f(\theta, n)$ the

	QR										OLS	
	$\tau 1$		$\tau 2$		$\tau 3$		$\tau 4$		$\tau 5$			
Parameter	value	P	value	P	value	P	value	P	value	P	value	P
Intercept	-1.66 (0.01)	0.00	-1.39 (0.01)	0.00	-1.09 (0.01)	0.00	-0.69 (0.01)	0.00	-0.24 (0.02)	0.00	-1.01 (0.01)	0.00
8DAF	0.10 (0.01)	0.00	0.16 (0.01)	0.00	0.24 (0.01)	0.00	0.29 (0.01)	0.00	0.33 (0.01)	0.00	0.23 (0.01)	0.00
9DAF	0.12 (0.01)	0.00	0.20 (0.01)	0.00	0.33 (0.01)	0.00	0.41 (0.01)	0.00	0.46 (0.01)	0.00	0.31 (0.01)	0.00
10DAF	0.25 (0.01)	0.00	0.43 (0.01)	0.00	0.65 (0.01)	0.00	0.82 (0.01)	0.00	0.93 (0.01)	0.00	0.62 (0.01)	0.00
11DAF	0.30 (0.01)	0.00	0.57 (0.01)	0.00	0.88 (0.01)	0.00	1.10 (0.01)	0.00	1.22 (0.01)	0.00	0.83 (0.01)	0.00
S1	-0.06 (0.01)	0.00	-0.13 (0.01)	0.00	-0.13 (0.01)	0.00	-0.17 (0.02)	0.00	-0.25 (0.02)	0.00	-0.15 (0.01)	0.00
S3	0.00 (0.01)	0.99	0.00 (0.01)	0.64	-0.01 (0.01)	0.28	-0.05 (0.02)	0.00	-0.11 (0.02)	0.00	-0.04 (0.01)	0.00
S4	0.04 (0.01)	0.00	0.05 (0.01)	0.00	0.07 (0.01)	0.00	0.07 (0.01)	0.00	-0.02 (0.02)	0.19	0.04 (0.01)	0.00
S1:S3	0.01 (0.02)	0.63	0.02 (0.02)	0.21	0.04 (0.01)	0.03	0.10 (0.02)	0.00	0.17 (0.03)	0.00	0.05 (0.01)	0.00
S1:S4	0.00 (0.02)	0.69	0.00 (0.02)	0.69	0.00 (0.01)	0.91	-0.01 (0.02)	0.57	0.04 (0.02)	0.10	0.00 (0.01)	0.99
S3:S4	-0.07 (0.02)	0.00	-0.14 (0.02)	0.00	-0.18 (0.01)	0.00	-0.22 (0.02)	0.00	-0.20 (0.03)	0.00	-0.15 (0.01)	0.00
S1:S3:S4	0.06 (0.03)	0.05	0.09 (0.03)	0.00	0.09 (0.01)	0.00	0.08 (0.03)	0.01	0.00 (0.04)	0.89	0.10 (0.02)	0.00

Table 1. Adjusted parameter estimated value (standard error), and their p-values P given for each estimator (QR with selected quantiles, and OLS) : QR = quantile regression, OLS = ordinary least square, $\tau 1$ = 0.1 quantile, $\tau 2$ = 0.25 quantile, $\tau 3$ = 0.5 quantile, $\tau 4$ = 0.75 quantile, $\tau 5$ = 0.9 quantile.

model used for each observation n.
Unlike OLS, QR utilizes the minimization of:

$$\widetilde{\theta}_\tau = argmin_\theta \; [\sum_{n:y^n \geq f(\theta,n)} \tau \left| y^n - f(\theta, n) \right| + \sum_{n:y^n < f(\theta,n)} (1-\tau) \left| y^n - f(\theta, n) \right|] \qquad (2)$$

for a given quantile τ. We used QR with five selected quantiles denoted $\tau 1$, $\tau 2$, $\tau 3$, $\tau 4$ and $\tau 5$ for respectively 0.1, 0.25, 0.5 (the median), 0.75, and 0.9 quantiles. All analyses were done using R project. The model was implemented in the function rq of the quantreg R-package for the QR estimations.

3. RESULTS AND DISCUSSION

Focusing on oleosin factors using the QR estimator (Table 1), the effect of S1 oleosin was statistically significant for all quantiles (p-value < 0.001) and with negative values of parameters decreasing with the increase of quantile (Figure 1). This reflects the contribution of S1 oleosin to the reduction of the volume of OB. Similarly, S3 oleosin factor had a significant effect of OB reduction, but only at higher OBs volume ($\tau 4$, $\tau 5$) and was not significant for lower volumes (p-value $= 0.99$, $\tau 1$; p-value = 0.64, $\tau 2$ and p-value $= 0.28$, $\tau 3$). A significant effect of reduction was also shown for S3 with the OLS. The effect of these two oleosins was clearly observed in the predictions sorted by genotype where the fitted values of OB volumes were always higher in genotypes lacking S1 and S3 oleosins (s1s3 and s1s3s4) compared to those where S1 and S3 oleosins were present (Figure 2). Unlike S1 and S3, S4 oleosin factor impact was significant with a positive effect (increase of OB volume) in all quantiles and within OLS estimator, except in $\tau 5$ (p-value $= 0.19$). It is noteworthy that the effect of increasing OB volume brought by the presence of S4 was smaller than the effect of reduction due to the presence of S1. This is shown by the low positive parameters values of S4 factor compared to the high negative parameters values of S1 factor (Figure 1). A significant synergistic effect of S3 and S4 interaction was shown, thus shedding light on the effect of interactions that also participate in determining OB volume. OB volume was reduced even more when S3 and S4 were both present. This interaction effect was visible both with the QR and the OLS estimators (p-values < 0.01). The other interaction effects were less significant for S1 and S3, except in $\tau 4$, and $\tau 5$. A null interaction between S1 and S4 was clearly shown both within the QR in low (p-value $= 0.69$, $\tau 1$) and high quantiles (p-value $= 0.57$, $\tau 4$) as well as with the OLS (p-value $= 0.99$). The triple interaction between S1, S3 and S4 was significant only in middle quantiles, and had no effect for higher volumes (p $= 0.89$, $\tau 5$).
The heterogeneous distribution of OBs requires the use of QR model in addition to OLS model since we were able to access more detailed effects of factors, while OLS model only indicated the differences on the central portion of the distribution. Particularly, QR model had the advantage of showing the significance of factors along quantiles. Our results show that the OB volume changes in function of time (between 7 and 11 DAF) and in function of the composition in oleosin proteins. The growth rate of the OB volume is shown to be reduced between 8 and 9 DAF, then increases between 9 and 10 DAF (Figure 2). While S1 and S3 oleosins have an effect on reducing the volume of OBs when they are present on the surface of OBs, S4 seems to be implicated on the increase of OB volumes except for

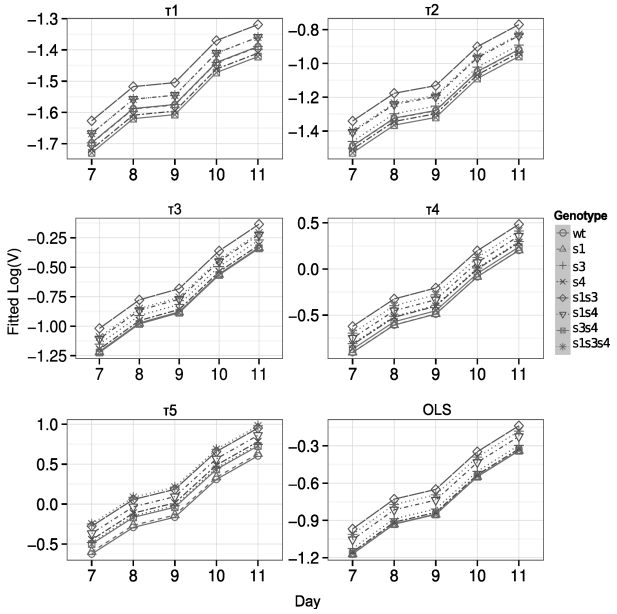


Figure 2. Predicted values of $Log_{10}(V)$ by date sorted by genotype for each estimator (QR with the selected quantiles and OLS) : Predicted values for wt and null oleosin mutant are deduced from combining the effect of each oleosin factor and their interactions respectively : wt (the effect of S1, S3, S4, S1:S3, S1:S4, and S1:S3:S4), s1 (effect of S3, S4, and S3:S4), s3 (effect of S1, S4, and S1:S4), s4 (effect of S1, S3, and S1:S3), s1s3 (effect of S4), s1s4 (effect of S3), s3s4 (effect of S1) and s1s3s4 (intercept).

higher ranges of volumes. One can also hypothesize that S4 could be involved in bringing OBs close enough so that their coalescence becomes easier. The null effect of S4 in higher volumes may be explained by the fact that an increased density of oleosins at the surface of OB prevents their coalescence when two OBs get closer. Furthermore, S1 and S4 do not interact and their effects seem to be only additive. Last, S3 and S4 present a high interaction effect of OB reduction. This may reflect an opposite action of S3 when S4 acts to increase OB volume.

4. CONCLUSION

The population of OBs in the cellular pool of *A. thaliana* embryos is heterogeneous. It is mainly composed of small OBs with a volume less than 1 μm^3 and few large OBs with a volume up to 20 μm^3. The growth mechanism of OBs is not fully understood, but many factors are suggested to be involved. Particularly, membrane composition may act on the dynamics of OBs including production, storage and mobilization of triacylglycerols. Furthermore, the membrane constituting elements follow themselves dynamical behaviors e.g. interaction and diffusion. In cells of plant seed, electrostatic and/or steric interactions may occur between oleosins on the surface of OBs. Moreover, the coalescence process causes an increase of the surface density. Based on these facts, one can suggest the probable variable effects of oleosins in function of OB size. Using OLS, we only focus on the effect of oleosins on the mean volume of OBs and disregard their variable effect due to the growth of OBs by coalescence. The QR model provides a better understanding of the effect of oleosins in different ranges of OB volumes. For example, OLS revealed that S3 oleosin reduced significantly the size of OBs (Table 1). However, QR showed that this effect was only significant on large OBs ($\tau 4$, $\tau 5$) and had no effect on small OBs, and this may be explained by the fact that S3 was only functional when reaching a certain surface coverage. The same behavior is observed with respect to S1 and S3 interactions (Table 1). A better understanding of the parameters controlling the growth of OBs may open new perspectives on oil storage and extraction techniques improvements. This statistical study will help in the design of a mechanistic model of OBs dynamics.

5. REFERENCES

[1] D. J. Murphy and I. Cummins, "Seed oil-bodies: Isolation, composition and role of oil-body apolipoproteins," *Phytochemistry*, vol. 28, no. 8, pp. 2063–2069, 1989.

[2] A. Huang, "Oleosins and oil bodies in seeds and other organs," *Plant Physiology*, vol. 110, no. 4, pp. 1055–1061, Jan. 1996.

[3] H. Yang, A. Galea, V. Sytnyk, and M. Crossley, "Controlling the size of lipid droplets: lipid and protein factors," *Current Opinion in Cell Biology*, vol. 24, no. 4, pp. 509–516, Aug. 2012.

[4] P. Jolivet, E. Roux, S. dAndrea, M. Davanture, L. Negroni, M. Zivy, and T. Chardot, "Protein composition of oil bodies in arabidopsis thaliana ecotype WS," *Plant Physiology and Biochemistry*, vol. 42, no. 6, pp. 501–509, June 2004.

[5] R. M. P. Siloto, K. Findlay, A. Lopez-Villalobos, E. C. Yeung, C. L. Nykiforuk, and M. M. Moloney, "The accumulation of oleosins determines the size of seed oilbodies in arabidopsis," *The Plant Cell Online*, vol. 18, no. 8, pp. 1961–1974, Jan. 2006.

[6] R. Koenker and G. Bassett, "Regression quantiles," *Econometrica*, vol. 46, no. 1, pp. 33, Jan. 1978.

[7] J. Levin, "For whom the reductions count: A quantile regression analysis of class size and peer effects on scholastic achievement," *Empirical Economics*, vol. 26, no. 1, pp. 221–246, 2001.

[8] D. Makowski, T. Dor, and H. Monod, "A new method to analyse relationships between yield components with boundary lines," *Agronomy for Sustainable Development*, vol. 27, no. 2, pp. 119–128, June 2007.

[9] J. Boulanger, C. Kervrann, P. Bouthemy, P. Elbau, J.-B. Sibarita, and J. Salamero, "Patch-based nonlocal functional for denoising fluorescence microscopy image sequences," *IEEE transactions on medical imaging*, vol. 29, no. 2, pp. 442–454, Feb. 2010, PMID: 19900849.

INFLUENCE OF CROSS SECTION SHAPE ON STEADY AND UNSTEADY FLOW THROUGH A CONSTRICTED CHANNEL

Bo Wu[1]*, Annemie Van Hirtum*[1] *and Xiaoyu Luo*[2]

[1]GIPSA-lab, UMR CNRS 5216, Grenoble University, France
[2]School of Mathematics and Statistics, University of Glasgow, Glasgow G12 8QW, UK
bo.wu@gipsa-lab.fr, annemie.vanhirtum@gipsa-lab.fr

ABSTRACT

Constricted channel flow at moderate Reynolds numbers is of physiological importance. The influence of the cross section shape on steady and unsteady flow through a constricted channel is assessed. Due to viscous effects the cross section shape can not be neglected since the pressure distribution depends on it. A flow model is proposed accounting for flow inertia, viscosity and the cross section shape in case of laminar steady and unsteady flow. Next, experiments are performed to characterise the influence of the cross section shape. Finally, the model outcome is validated on experimental data.

1. INTRODUCTION

Pressure driven channel flow is associated with physiological flows for which constricted channel portions occur either naturally or due to a pathology. Well known examples are airflow through the human airways (human speech production, asthma, obstructive sleep apnea) or blood flow through a stenosis.

Consequently, efforts are made to model pressure driven flow through constricted channels in order to understand the mechanisms involved and to develop aiding tools for health care workers. Due to the complexity of the human respiratory and cardiovascular system, most studies severely simplify the physiological reality in order to limit the number of physiological and physical parameters. Such a simplification enhances understanding and facilitates implementation and experimental validation.

In general, simplifications of the flow model are based on a non dimensional analysis of the governing Navier-Stokes equations [1]. Accounting for typical values of physiological, geometrical and flow characteristics result in non dimensional numbers which allows one to assume the flow as incompressible, laminar, quasi one or two dimensional and quasi steady [2, 3]. Therefore, quasi-one dimensional or two dimensional (2D) flow models derived from boundary layer theory have proven to capture the underlying physics and therefore to mimic and predict ongoing phenomena at a low computational cost [2].

Nevertheless, the assumption of a 1D or 2D geometry implies that details of the cross section shape perpendicular to the streamwise flow direction x are neglected. Viscous effects, which are important at low Reynolds numbers, are known to depend on the cross section shape [1, 3]. The aim of the present paper is therefore to propose a flow model capable to account for flow inertia, viscosity as well as for the cross section shape in case of steady and unsteady flow. Experimental data are presented in order to assess the influence of cross section shape for steady and unsteady flow. The model outcome is validated.

2. CROSS SECTION SHAPES

The geometry is fully defined by the cross section shape and the streamwise area variation $A(x)$. In order to allow the use of the cross section shapes in quasi-analytical models only shapes for which the main geometrical parameters can be expressed analytically are assessed: rectangle (re), circle (cl), ellipse (el), eccentric annulus (ea), half moon (hm), circular sector (cs), equilateral triangle (tr) and limacon (lm) [3]. The cross section shapes used during experiments are illustrated in Fig. 1. The shapes are, although a severe approximation, relevant to describe the cross section shapes in case of normal and/or pathological geometrical conditions. The cross section is positioned in the (y, z) plane where y denotes the spanwise and z the transverse direction. The cross section area is uniform and yields $A = 0.79\text{cm}^2$. Corresponding values for the hydraulic diameter D ($D = \frac{4A}{P}$ with perimeter P) are given in Table 1.

Table 1. Overview geometrical parameters.

	cl	re	el	sq	tr	ntr	scs	lcs
D[mm]	10	6.6	6.7	8.9	7.8	7.0	7.2	8.4

$A = 0.79\text{cm}^2$

3. MODELLING

The streamwise area variation consists of a uniform constriction, with fixed length $L = 25$mm and varying cross section shape, which is inserted in a uniform tube of internal diameter 25mm as schematically depicted in Fig. 2. As the abrupt expansion is characterized by a sharp trailing edge, the streamwise position of flow separation x_s is

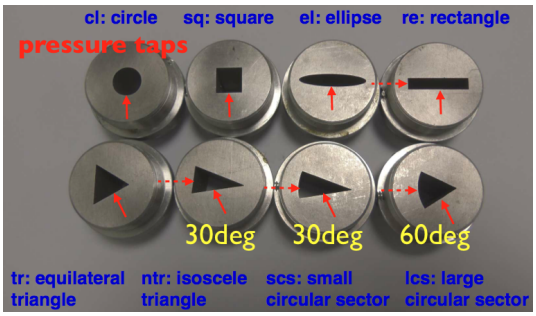


Figure 1. Experimentally assessed cross section shapes with pressure taps P_1 (full arrow) and P_2 (dashed arrow).

at the constriction end ($x_s = x_3$). The pressure downstream from the flow separation point is assumed to be zero so that $P_d = 0$ and the model outcome remains constant for $x \geq x_s$. Consequently, imposing the upstream pressure P_0 is equivalent to imposing the driving pressure gradient $P_0 - P_d$.

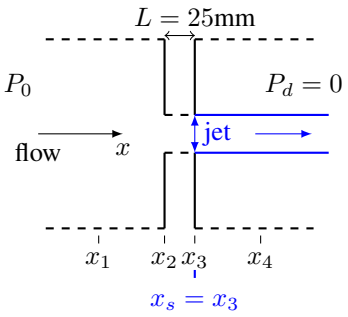


Figure 2. Flow through an abrupt expansion.

For a given fluid and under assumption of a laminar and incompressible flow the streamwise momentum equation of the governing Navier-Stokes equations is simplified using additional assumptions. With driving pressure gradient dP/dx, bulk velocity $\overline{U}$, cross section area A, volume flow rate Q, velocity $u(x, y, z)$, kinematic viscosity ν (1.5×10^{-5}m^2/s for air) and density ρ (1.2kg/m^3 for air) the following flow models are assessed:

- Applying conservation of volume flow rate $\frac{dQ}{dx} = 0$, the following simplified flow model accounts for both viscosity and flow inertia and depends therefore on both shape and area of cross section:

$$\frac{d\overline{U}}{dt} - \frac{Q^2}{A^3}\frac{dA}{dx} = -\frac{1}{\rho}\frac{dP}{dx} + \nu\left(\frac{\partial^2 u}{\partial y^2} + \frac{\partial^2 u}{\partial z^2}\right) \quad (1)$$

  Making a 2D assumption allows to drop the first term within brackets resulting in the quasi-one dimensional flow model, which is further labelled Bernoulli-Poiseuille flow (BP) [2]. The first term at the left handside accounts for flow unsteadiness. Note that in the current paper, unsteadiness is due to varying the upstream flow conditions, *i.e.* $P_0(t)$, whereas the cross section area is time independent.

- Thwaites' method for 2D and axisymmetrical flow solving steady integral momentum equation [1, 4]:

$$\frac{d\delta_2}{dx} + \left(2 + \frac{\delta_1}{\delta_2}\right)\frac{\delta_2}{U_e(x)}\frac{dU_e(x)}{dx} = \frac{\tau}{\rho U_e^2(x)} \quad (2)$$

  with flow velocity outside the boundary layer U_e and wall shear stress τ. Displacement thickness $\delta_1(x)$ and momentum thickness $\delta_2(x)$ are defined:

$$\delta_1(x) = \int_0^\infty R^k(x)\left(1 - \frac{u(x,y)}{U_e(x)}\right)dy$$
$$\delta_2(x) = \int_0^\infty R^k(x)\frac{u(x,y)}{U_e(x)}\left(1 - \frac{u(x,y)}{U_e(x)}\right)dy$$
$$\text{flow index } k = \begin{cases} 0 & \text{2D flow} \\ 1 & \text{axisymmetrical flow} \end{cases} \quad (3)$$

For uniform geometries and applying the no-slip boundary condition $u = 0$ on the channel walls, Eq. 1 can be rewritten as a classical Dirichlet problem which can be solved analytically for simple cross section shapes, such as the geometries shown in Fig. 1. Therefore, exact solutions are obtained for: local velocity $u(x, y, z)$, local pressure $p(x)$, wall shear stress τ and volume flow rate Q [3]. With these notations bulk Reynolds number $Re = \frac{QD}{\nu A}$ and Strouhal number $Sr = \frac{f_0 DA}{Q}$ are defined using hydraulic diameter D and characteristic frequency f_0.

4. EXPERIMENT: SETUP AND CONDITIONS

The flow facility, illustrated in Fig. 3, consists of an air compressor (Atlas Copco GA7), followed by a pressure regulator (Norgren type 11-818-987) providing an airflow at constant pressure. The volume flow rate is controlled by a manual valve placed downstream the regulator and measured by a thermal mass flow meter (model 4043 TSI) with an accuracy of 2% of its reading. To homogenize the flow, a settling chamber is used with volume $0.25\times0.3\times0.35$m^3 to which a series of 3 perforated plates with holes of diameter 1.5mm are added. The walls of the settling chamber are tapered with acoustic foam (SE50-AL-ML Elastomeres Solutions) in order to avoid acoustic resonances. The influence of the cross section shape on the flow development is assessed experimentally by adding a constriction with different cross section shape (Fig. 1), fixed area $A = 0.79$cm^2 and fixed length $L = 25$mm to a uniform circular tube, with diameter 25mm, mounted to the settling chamber by means of a converging nozzle. An acoustic pressure driver unit (Ku 916T) upstream the constriction is used to create unsteady flow. Pressure sensors (Kulite XCS-093) are positioned in pressure tap of diameter 0.5mm upstream (P_0) and in the middle (P_1) of the constricted portion. Volume flow rate Q is sampled at 500Hz. Pressure sensors P and microphones M are sampled at 24kHz. Statistical quantities, such as mean and root mean square (rms) pressure, are derived on 5s of signal corresponding to 120000 samples.

Steady flow is studied for volume flow rates within the range $0 \leq Q \leq 200$l/min. The increment is 5l/min for $Q \leq 80$l/min, 90l/min and 25l/min for $Q \geq 100$l/min. Unsteady flow is assessed for $Q = 5$, 20 and 150l/min with driving frequency $f_0 = 500$Hz.

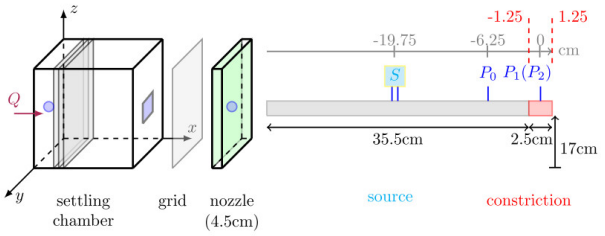


Figure 3. Illustration of the experimental setup.

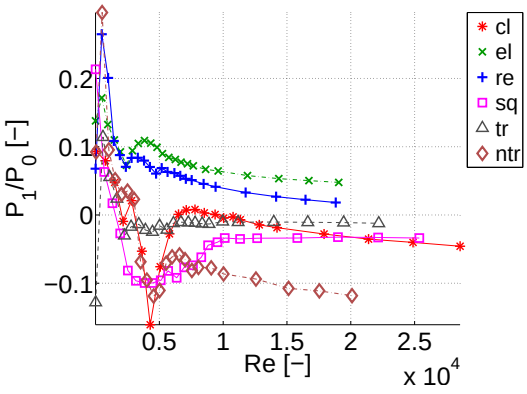
5. RESULTS

5.1. Experimental data: steady flow

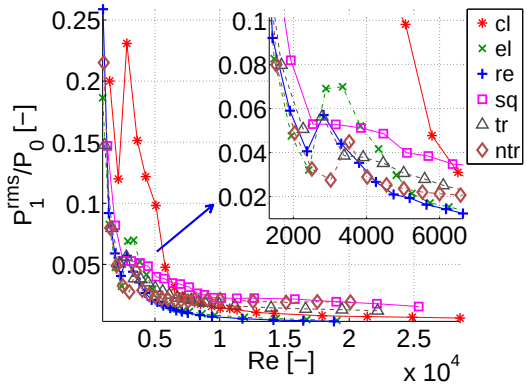
The pressure distribution in a constricted channel with fixed streamwise area is experimentally assessed for steady flow. Mean and rms pressures within the constriction normalised by the mean upstream pressure P_0 are illustrated in Fig. 4. The mean and rms values vary up to $\approx 25\%$ of P_0, which confirms the need to take into the cross section shape. For all cross section shapes, the general decreasing tendency in case of both mean and rms pressure values within the constriction is observed to change in the range $2000 < Re < 4000$. Indeed, in this Reynolds number range an increase is observed so that a minimum and peak values occur for both the normalized mean and rms pressure for all assessed cross sections. This range of Reynolds numbers is likely to be associated with the passage of vortices generated at the sharp constriction inlet or at the outlet. Either way, vortex generation and interaction is likely to cause a transition from laminar to turbulent flow. Further research is required to fully determine the flow dynamics.

5.2. Experimental data: unsteady flow

Measured pressures $P_0(t)$ and $P_1(t)$ in case of a circular and elliptic cross section shape are illustrated in Fig. 5. The unsteady oscillatory flow $P_0(t)$ with period $T = 1/f_0$ illustrates flow for $Sr \approx 1$ and $Sr < 1$. As for steady flow the mean pressure value within the constriction varies as function of the cross section shape, *e.g.* the ratio observed for the elliptic section is greater than the one observed for the circular cross section. In addition, the amplitude of the pressure in the constriction around its mean value, $P_1(t) - \overline{P_1}(t)$, observed for the elliptic section is greater than the one observed for the circular cross section. Moreover, a phase difference between the upstream pressure P_0 and constriction pressure P_1 is observed, which is seen to depend on both Reynolds number and cross section shape as summarised in Table 2. As for the steady flow, further research is needed to fully determine the flow dynamics.



(a) P_1/P_0



(b) P_1^{rms}/P_0

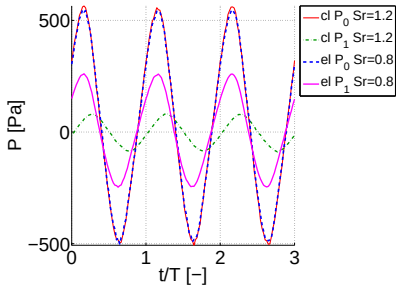
Figure 4. Normalised mean and rms pressures within the constriction for steady flow.

Table 2. Normalized phase difference of $P_0(t)$ and $P_1(t)$.

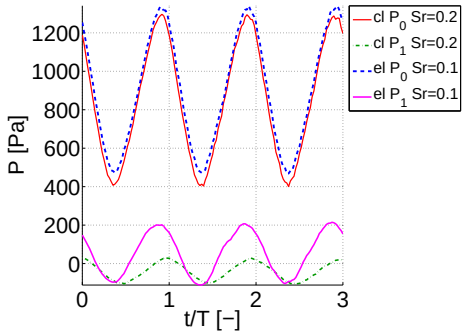
Q[l/min]	$\frac{t}{T}[-]$					
	cl	re	el	sq	tr	ntr
5	0.08	0	0.02	0.98	0.98	0.02
20	0.10	0.02	0.02	0.98	0.98	0.04
150	0.04	0.02	0.94	0.90	0.94	0.92

5.3. Model validation

Fig. 6(a) illustrates normalized modeled and experimental pressures within the constriction, P_1/P_0, for steady flow. For all assessed geometries the variation of the normalised model outcome accounting for the cross section shape is within 5%. Moreover, since a uniform constriction is considered all flow models predict positive pressures at position P_1 so that the negative pressures occurring for all assessed cross section shapes, except the rectangle, can not be predicted with the used flow models. In addition, the model outcome as a function of increasing Reynolds number results in monotonously decreasing values of P_1/P_0 so that more complex flow phenomena such as observed in the range $2000 < Re < 4000$ can not be captured with the assessed flow models. Nevertheless, the discrepancy



(a) $Q = 20$l/min, $Re \approx 2000$



(b) $Q = 150$l/min, $Re \approx 1.4 \times 10^4$

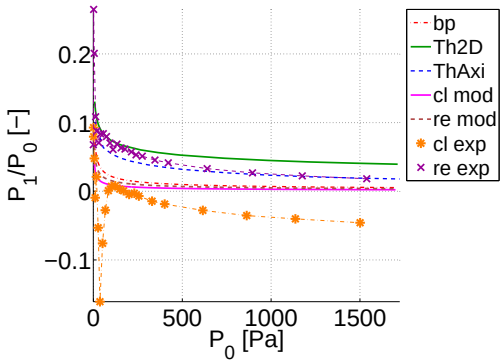
Figure 5. Measured pressures P_0 and P_1 for a circular (cl) and elliptic (el) cross section.

between modeled and measured values is quantified.

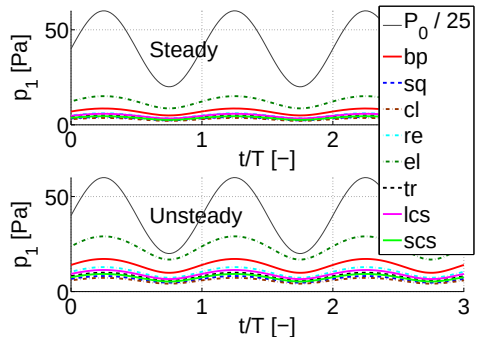
In addition to the model accounting for the exact cross section shape (mod), Fig. 6(a) depicts Thwaites' 2D and axismmetrical solution (Th) and 2D Bernoulli-Poiseuille (bp). For a rectangular cross section shape, the model output accounting for the cross section shape and Bernoulli-Poiseuille 2D flow coincides and result in an overall overestimation of the pressure drop within 5% for $P_0 > 300$Pa and within 10% for $P_0 < 300$Pa. Thwaites 2D solution severely underestimates the pressure drop, $\approx 5\%$ for all upstream pressures. In case of a circular cross section, the axisymmetrical Thwaites solution severely overestimates measured values, $> 10\%$ for all upstream pressures. The model accounting for the cross section shape underestimates the pressure drop within 5% for $P_0 > 300$Pa and fails for $P_0 < 300$Pa. In general, accounting for the cross section shape and 2D Bernoulli-Poiseuille both result in a model accuracy of $< 5\%$ for $P_0 > 300$ and of $<5\%$ up to $< 20\%$ for $P_0 < 300$ depending on the cross section shape. Fig. 6(b) illustrates the scaled upstream pressure and corresponding model outcome while accounting for the cross section. The influence of the unsteady term is apparent.

6. CONCLUSION

The influence of the cross section shape on flow through a constricted channel is shown for experimental and modeled data. While the flow model is not able to account for



(a) circular and rectangular shape



(b) mod: all cross section shapes

Figure 6. Modelled (mod, Th, bp) and experimental (exp) values of P_1/P_0 for steady (a,b) and unsteady (b) flow.

complex flow dynamics such as vortex generation, interaction or turbulence, it does provide a 5% accurate pressure prediction for $P_0 > 300$Pa. Consequently, it is of use to improve flow models used in mathematical or physical models of physiological flow driven events such as fluid-structure interactions in the upper airways, *(*etc.)

7. ACKNOWLEDGMENTS

Support from the French Ministry of Education and Research and Royal Society (UK) is acknowledged.

8. REFERENCES

[1] F. M. White, *Viscous Fluid Flow*, McGraw-Hill, New York, 2nd edition, 1991.

[2] J. Cisonni, A. Van Hirtum, X. Pelorson, and J. Willems, "Theoretical simulation and experimental validation of inverse quasi one-dimensional steady and unsteady glottal flow models," *J. Acoust. Soc. Am.*, vol. 124, pp. 535–545, 2008.

[3] B. Wu, A. Van Hirtum, and X. Y. Luo, "Pressure driven steady flow in constricted channels of different cross section shapes," *Int. J. Applied Mechanics*, Accepted.

[4] H. Schlichting and K. Gersten, *Boundary Layer Theory*, Springer, Berlin, 8th edition, 2000.

ABSTRACTS

MULTIRESOLUTION MIXTURE MODELS USING BAYESIAN NETWORKS

Prem Raj Adhikari[1,2] *and Jaakko Hollmén*[1,2]

[1]Helsinki Institute for Information Technology,
[2]Department of Information and Computer Science,
Aalto University School of Science,
PO Box 15400, FI-00076 Aalto, Espoo, Finland
prem.adhikari@aalto.fi, jaakko.hollmen@aalto.fi

ABSTRACT

Multiresolution data arises when the same phenomenon or the data generating process is measured at different levels of precision. If a system is measured exhaustively in detail, it produces data in fine resolution where as if the same system is measured in-comprehensively, the data is produced in coarse resolution. This phenomenon is present in many systems because the newer generation technology can measure the finer units of the system producing data in fine resolution. In contrast, the older generation technology produces the data in coarse resolution as the older generation technology can not measure the finer units of the data. In this scenario, when the same data generating process generates data in coarse and the fine resolution, often a feature in coarse resolution is the amalgamation of multiple features in the fine resolution. Therefore, the dataset can be represented in the form of a tree where the features of the data in the coarse resolution forms the root of the tree and the features of the data in the fine resolution forms the branches and the leaves of the tree. In a typical case of multiresolution data represented in trees, the ancestors are the features of the data in coarse resolution where as the children are the features of the data in the fine resolution. The number of nodes arising from a root node to its branches determines the number of different features in fine resolution originating from a single feature in the coarse resolution. These relationships between the features in different resolutions of the data ascertained from the knowledge of the application domain. We exploit such structures of multiresolution data in to propose a multiresolution mixture model in this contribution.

Learning from multiresolution data is of paramount importance because machine learning algorithms are always data hungry and single resolution datasets are always constraint by high data dimensionality coupled with the lack of large number of data samples. Mixture models have been versatile in modelling diverse phenomenon for over a century. However, the mixture model in its general form can only model the data in a single resolution i.e. data having the same dimensionality. Currently, only mixture modelling solution to multiresolution data is to model the different resolutions separately and at best compare them. In such scenario, a multiresolution mixture model can be an antidote to such compounding problems of multiresolution data and also exploiting the benefits of mixture models. We learn a single multiresolution model unlike our previous approach [1] where we generate a model each for each resolution of the data although the model in single resolution absorbed the information in other resolution of data because of repeated merging of mixture components. Learning mixture model involves: determining the number of mixture components and inferring the parameters of each mixture component [2]. In this contribution, we represent multiresolution data in the form of a collection of probabilistic graphical models and the components of the mixture model are graphical models itself. However, learning a structured multiresolution mixture model is difficult because some resolutions can be missing from the data. Our research considers this problem and proposes a probabilistic approach using Bayesian networks to learn a multiresolution mixture model even when data in some resolutions are missing. We experimented our algorithm on a multiresolution artificial dataset generated such that it mimics the properties of a multiresolution chromosomal aberration dataset. Similarly, we also experimented on a real world multiresolution chromosomal aberration dataset. The results in both the cases show that multiresolution mixture models outperform single resolution models.

1. REFERENCES

[1] P. R. Adhikari and J. Hollmén, "Multiresolution Mixture Modeling using Merging of Mixture Components," in *Proceedings of the Fourth Asian Conference on Machine Learning*, S. C. H. Hoi and W. Buntine, Eds. 2012, vol. 25 of *ACML'12, Singapore*, pp. 17–32, JMLR Workshop and Conference Proceedings.

[2] P. R. Adhikari and J. Hollmén, "Fast Progressive Training of Mixture Models for Model Selection," in *Proceedings of Fifteenth International Conference on Discovery Science (DS 2012)*, J.-G. Ganascia, P. Lenca, and J.-M. Petit, Eds. October 2012, vol. 7569 of *Lecture Notes in Artificial Intelligence*, pp. 194–208, Springer-Verlag.

INTEGRATIVE SEQUENCING REVEALS ALTERATIONS IN UNTREATED AND CASTRATION RESISTANT PROSTATE CANCER

Matti Annala[1,2], Kati Waltering[1,2]*, Antti Ylipää[1,2]*, Kimmo Kartasalo[1,2], Kirsi Tuppurainen[2], Leena Latonen[2], Simo-Pekka Leppänen[1,2], Serdar Karakurt[2], Outi Saramäki[2], Mauro Scaravilli[2], Janne Seppälä[1,2], Hanna Rauhala[2], Olli Yli-Harja[1,3], Robert Vessella[4], Teuvo Tammela[5], Wei Zhang[6], Tapio Visakorpi[2,3] and Matti Nykter[1,2,3]*

[1]Department of Signal Processing, Tampere University of Technology, Tampere, Finland
[2]Institute of Biomedical Technology, University of Tampere, Tampere, Finland
[3]BioMediTech, Tampere, Finland
[4]Department of Urology, University of Washington, Seattle, WA, USA
[5]Department of Urology, Tampere University Hospital and Medical School, Tampere, Finland
[6]Department of Pathology, University of Texas MD Anderson Cancer Center, Houston, TX, USA
P.O. Box 553, FI-33101 Tampere, Finland,
matti.annala@tut.fi

* These authors contributed equally.

ABSTRACT

Castration resistant prostate cancer (CRPC) is the third most common cause of male cancer death in developed countries. Previous sequencing studies have focused on specific aspects of prostate cancer biology. Here we report the integrative sequencing of genomic, transcriptomic and DNA methylation changes in 28 untreated prostate cancers and 13 CRPCs. AR, TGF-β and WNT signaling pathways were recurrently altered in CRPC. We identified two new functionally relevant fusion genes, TMPRSS2-SKIL and DOT1L-HES6. Fusion analysis in an independent cohort validated SKIL's role as a recurrent 3' fusion partner and oncogene in prostate cancer. The HES6 fusion was found in an AR-negative CRPC, and its overexpression in vitro led to androgen independent growth. Transcriptome assembly uncovered 128 previously unannotated prostate cancer associated transcripts, including the ERG regulated transcript TPCAT-10-36067 whose knockdown had a dramatic effect on prostate cancer cell growth and apoptosis. Our data provides a unique resource for the prostate cancer research community and presents new opportunities for therapeutic intervention.

CHROMOTHRIPSIS-LIKE PATTERNS ARE RECURRING BUT HETEROGENEOUSLY DISTRIBUTED FEATURES IN A SURVEY OF 22,347 CANCER SAMPLES

Haoyang Cai[1,2], Nitin Kumar[1,2], Homayoun C. Bagheri[3], Christian von Mering[1,2], Mark D. Robinson[1,2] and Michael Baudis[1,2]

[1]Institute of Molecular Life Sciences, University of Zurich, Zurich, Switzerland
[2]Swiss Institute of Bioinformatics, University of Zurich, Zurich, Switzerland
[3]Institute of Evolutionary Biology and Environmental Studies, University of Zurich, Zurich, Switzerland
haoyang.cai@imls.uzh.ch

ABSTRACT

Introduction: Chromothripsis is a newly discovered type of genomic rearrangement, characterized by locally clustered copy number aberrations. It has been proposed that it may arise during a single genome-shattering event. In this model, contiguous chromosomal regions are fragmented into many pieces. Supposedly, these segments are then randomly fused together by the cell's DNA repair machinery. This "shattering" and aberrant repair of a multitude of DNA fragments could provide an alternative oncogenetic route, in contrast to the step-by-step paradigm of cancer development. However, the underlying mechanisms and their specific impact on tumorigenesis are still poorly understood.

Results: Here, we identified chromothripsis-like genome patterns in 918 cancer samples, from a dataset of more than 22,000 oncogenomic arrays covering 132 cancer types. Fragmentation hotspots were found to be located on chromosome 8, 11, 12 and 17. The uneven distribution of chromothripsis along the tumor genomes may reveal associations between tumor type related cancer genes and molecular mechanisms behind chromosome-shattering events. Among the various cancer types, soft-tissue tumors exhibited particularly high CTLP frequencies. Genomic context analysis revealed that chromothriptic rearrangements frequently occurred in genomes that additional harbored multiple copy number aberrations (CNAs). Therefore, for those frequent cases exhibiting additional non-chromothripsis CNA events, their possible contribution to oncogenesis has to be considered when modeling the role of chromothripsis in cancer development. Moreover, an investigation into the affected chromosomal regions showed a large proportion of arm-level pulverization and telomere related events, which would support breakage-fusion-bridge cycles as one of the potential underlying mechanisms. We also report evidence that this catastrophic event may be correlated with patient age, stage and a lower survival rate.

Conclusion: Chromothripsis-like patterns represent a striking feature occurring in a limited set of cancer genomes, and can reliably be detected using biostatistical methods. The observed patterns may reflect on heterogenous biological phenomena beyond a single class of "chromothripsis" events, and probably vary in their specific impact on oncogenesis. Fragmentation hotspots derived from our large-scale data set may promote the detection of markers involved in chromothriptic rearrangements, or may be used for assigning disease related effects to a chroothripsis induced genomic events.

GPU-POWERED SENSITIVITY ANALYSIS AND PARAMETER ESTIMATION OF A REACTION-BASED MODEL OF THE POST REPLICATION REPAIR PATHWAY IN YEAST

Paolo Cazzaniga[1], Riccardo Colombo[2], Marco S. Nobile[2], Dario Pescini[3], Giancarlo Mauri[3]and Daniela Besozzi[4]

[1]Università di Bergamo, Dipartimento di Scienze Umane e Sociali, Piazzale S. Agostino 2, 24129 Bergamo, Italy
[2]Università di Milano-Bicocca, Dipartimento di Informatica, Sistemistica e Comunicazione, Viale Sarca 336, 20126 Milano, Italy
[3]Università di Milano-Bicocca, Dipartimento di Statistica e Metodi Quantitativi, Via Bicocca degli Arcimboldi 8, 20126 Milano, Italy
[4]Università di Milano, Dipartimento di Informatica, Via Comelico 39, 20135 Milano, Italy
paolo.cazzaniga@unibg.it

ABSTRACT

Introduction: UV radiation induces DNA lesions that must be processed by the Post-Replication Repair (PRR) pathway in order to allow the completion of genome replication. PRR is triggered by ubiquitylation, a post-translational modification of the PCNA protein; in particular, mono-ubiquitylation of PCNA activates the mutagenic Translesion DNA Synthesis, while PCNA poly-ubiquitylation directs PRR to the non mutagenic Template Switching. Through a Systems Biology approach we previously developed a reaction-based model of the PRR in yeast [Amara et al., BMC Syst. Biol. 7:2013], investigated at different UV radiation doses, which correctly reproduces "in vivo" experimental PCNA ubiquitylation dynamics at low acute UV doses.

The reaction constants of the model were manually tuned exploiting the experimental time-courses of mono- and poly-ubiquitylated PCNA isoforms. The aim of this work is to develop a framework to automatically identify a plausible set of model parameters, by integrating sensitivity analysis (SA) with parameter estimation (PE) methods. In order to reduce the computational burden due to the large number of independent simulations required by SA and PE, we used cupSODA [Nobile et al., PaCT2013, accepted], our parallel GPU-based ODE integrator.

Results: We first analysed the PRR model by means of the improved Elementary Effects SA method [Campolongo et al., Comp. Phys. Comm. 182:2011], to measure the influence of each kinetic parameter on the dynamics of mono-, di- and tri- ubiquitylated PCNA isoforms.

SA results suggest that the most sensible reactions are those related to the formation of poly-ubiquitylated PCNA isoforms, whose sensitivity is mainly due to the contribution of the di-ubiquitylated isoforms. Moreover, these results enlighten that the sensitivity of each reaction is strongly influenced by the other kinetic parameters, underlying the system's non-linearity.

Exploiting the results of the SA, PE was then performed by means of a GPU-based multi-swarm version of the Particle Swarm Optimizer [Nobile et al., LNCS 7246:2012], where each swarm can be assigned to a specific experimental condition tested "in vivo". Thanks to its design, this methodology allows the simultaneous analysis of the behaviour of the PRR model at both low and high acute UV doses.

Conclusions: Our GPU-powered framework achieved a relevant speedup of the simulations required for SA and PE (compared to the sequential implementation), thus widening the analysed parameters space. This allowed a deeper and precise investigation of the PRR model, and suggested biological insights related to the control points of the PCNA ubiquitylation pathway.

THE DYNAMICS OF THE GENETIC REPRESSILATOR FOR VARYING TEMPERATURES

Jerome G. Chandraseelan, Samuel M.D. Oliveira and Andre S. Ribeiro

Laboratory of Biosystem Dynamics,
Department of Signal Processing, Tampere University of Technology,
P.O. Box 553, FI-33101 Tampere, Finland,
jerome.chandraseelan@tut.fi, samuel.matosdiaoliveira@tut.fi, andre.ribeiro@tut.fi

ABSTRACT

Genes form networks of interactions capable of performing complex functions that are not possible by single genes. The resulting Gene Regulatory Networks (GRN) is more robust to fluctuations and adaptable to a multitude of environments. Organisms make use of these GRN to perform a variety of processes such as counting time, responding to changes in the environment and regulating processes related to, for example, maintenance of the homeostasis of the metabolism or the developmental program. To better understand the regulatory mechanisms of GRNs, synthetic circuits with defined components have been developed. The Repressilator is likely the best known synthetic genetic circuit. It comprises three promoters connected in a negative feedback loop such that the activity of one gene in the circuit represses the action of the subsequent one in the loop. The resultant oscillation in protein numbers is read through the expression of a green fluorescent protein (GFP) reporter under the control of one of the promoters in the network. Here, we present measurements of the dynamics of this circuit at different temperatures, showing that the mean period is highly dependent on this environmental factor. The results aim to assist in developing artificial circuits that are robust to environmental changes by identifying the causes for the observed behavior modifications.

RANDOM BOOLEAN NETWORK BASES SIMULATION OF CELLULAR DIFFERENTIATION PROCESSES IN HUMAN IMMUNE SYSTEM

Mattia Cinelli[1] and Csaba Ortutay [2]

[1]Erasmus student from University of Bologna,
[2]Postdoctoral Fellow, Adjunct professor, University of Tampere,
Institute of Biomedical Technology, University of Tampere, Tampere, Finland,
mattia.cinelli@uta.fi
csaba.ortutay@uta.fi

ABSTRACT

Introduction:
Random Boolean Networks (RBNs) are computational tools to simulate the behavior of complex process. The investigation of cellular differentiation processes in human system applying RBNs is the topic and main aim of this project. Methodologies and tools, previously developed, are used to identify gene network relevant for B and T cells differentiation and these now are simulated and studied. In this project are proposing new methods and techniques in order to apply RBN system ad produce more reasonable results.

Methods:
The whole project is based on three foundations:
- Network decomposition: A method developed previously allows getting for the most important elements of B and T cell specific immune system. Based on correlation of gene expression changes. This is the best way for study the whole system as sum of its parts.
- Collection and analysis of row expression data: Microarray data are used for get Pearson's correlation coefficient between each gene pairs represented in the immune network.
-Software for Boolean network simulation: Is developed from the generation and study of attractors. The main difference between our program and earlier approaches is that for each nodes do not use the basic Boolean functions (and/or), but only one function with a random noise component.

Results:
The simulation software was used to generate the states of the B cell specific gene influence network. The attractor study provides clusters of stable states which correspond to B cell sub-types according to data from Gene Ontologies or from list of known clusters of differentiation (CD) markers from the literature.

Conclusion:
The study of complex biological system is a hard field of research and we suggest a novel approach. The results got are undoubtedly encouraging; in the simulation of B cell differentiation. Now the simulator is a gene basic method but we plan to implement it adding also the protein-protein interaction and reaction kinetics as the next step.

PARAMETER ESTIMATION FOR JAK-STAT MODEL VIA SENSITIVITY ANALYSIS

Krzysztof Fujarewicz and Krzysztof Łakomiec

Institute of Automatic Control, Silesian University of Technology,
Akademicka 16, 44-100 Gliwice, Poland,
krzysztof.fujarewicz@polsl.pl

ABSTRACT

Cell signaling pathways are usually modeled as systems of ordinary differential equations (ODEs). The problem of parameter estimation for given experimental data is usually solved by numerical minimization of an objective function which is defined as a measure of differences between measurements and model solution. The information about the gradient of the objective function with respect to the model parameters may significantly speed up the estimation process. The gradient can be achieved by using sensitivity analysis [Fujarewicz et al. 2007].

The sensitivity analysis plays very important and useful role during investigation and modeling of dynamical systems. The sensitivity analysis of dynamic systems answers the question how changes of model parameters affects the model solution. The answer to this question can be useful in solving of many tasks, such as: estimation of model parameters, design of experiments, or the optimization of the structure of the model. Typically, the sensitivity functions with respect to model's parameters are calculated but it is possible to perform the sensitivity analysis with respect to initial conditions or signals stimulating the system.

There are several practical approaches to determine the sensitivity functions, which can be divided into three groups: (i) finite difference approximation, which uses explicitly the difference quotient formula, (ii) forward sensitivity analysis, where so called sensitivity model (tangent-linearized) for the variations of signals is built and simulated, and (iii) adjoint sensitivity analysis, where so called adjoint system is built and simulated. Especially the third approach is very useful from practical point of view, because it minimizes the computational effort required to all sensitivity functions calculation.

One of the proposed in the literature mathematical model of the JAK-STAT signaling pathway is described by means of a set of delayed differential equations (DDE) [Timmer et al. 2004]. For the parameter estimation purposes the forward sensitivity analysis may be used [Loxton 2010]. Instead of this we present the way to perform the adjoint sensitivity analysis of DDEs. The method is structural because it assumes that the system is presented in a structural form: as a block diagram. It simplifies the rules for the adjoint system creation and may be treated as a special case of so called automatic differentiation. The results of the sensitivity analysis for JAK-STAT signaling pathway and its application of to parameter estimation are presented.

This work was supported by the Polish National Science Center under grant UMO-2012/05/B/ST6/03472.

Literature

Fujarewicz K., Kimmel M., Lipniacki T., Świerniak A.: Adjoint systems for models of cell signalling pathways and their application to parameter fitting, IEEE/ACM Transactions on Computational Biology and Bioinformatics, 4(3), p. 322-335, 2007.

Loxton R.,Teo K.L.,Rehbock V., An optimization approach to state-delay identification, IEEE Transactions on Automatic Control 55, 2113-2119, 2010.

Rihan F.A.: Sensitivity analysis for dynamic systems with time-lags, Journal of Computational and Applied Mathematics, 151(2), p. 445-462, 2003.

Timmer J., Müller T. G., Swameye I., Sandra O., & Klingmüller U.: Modeling the nonlinear dynamics of cellular signal transduction, International Journal of Bifurcation and Chaos, 14(06), 2069-2079, 2004.

CELL DIVISION ASYMMETRIES IN *ESCHERICHIA COLI* WHEN IN SUB-OPTIMAL CONDITIONS.

Abhishekh Gupta [1], Jason Lloyd-Price [1], Meenakshisundaram Kandhavelu [1], Samuel M.D. Oliveira [1] and Andre S. Ribeiro [1]

[1]Laboratory of Biosystem Dynamics, Computational Systems Biology Research Group,
Department of Signal Processing, Tampere University of Technology,
P.O. Box 553, FI-33101 Tampere, Finland,
abhishekh.gupta@tut.fi

ABSTRACT

Cell division in *Escherichia coli*, under optimal conditions, is generally morphologically symmetric. However, recent evidence has shown that this organism preferentially segregates unwanted substances to the older pole. This bias in the segregation to the poles has been shown to weaken with decreasing temperature and in minimal media. It is unknown whether there is any relationship between this asymmetry and other morphological asymmetries in these organisms.

Here, we investigate possible relationships between the asymmetric segregation of unwanted substances in *E. coli* and morphological asymmetries in the process of cell division. For this, first we characterize, in different environmental conditions, the variance in the sizes of the daughter cells immediately following division. Next, by tracking fluorescently-tagged aggregates with single-molecule sensitivity, we quantify the asymmetries in partitioning of MS2-GFP-RNA aggregates in division. We then investigate the qualitative relationship between this and the asymmetries in size.

We find that in optimal growth conditions (LB media at 37°C), divisions are more morphologically symmetric; though a strong asymmetry in partitioning of unwanted aggregates is observed. In this condition, these two quantities are not correlated. As conditions 'worsen', due to decreased temperature down to 24°C or due to being in minimal media (M63), the divisions became less symmetric, in that the variance in relative daughter sizes increased. Also, the shape, as measured by the roundness of the cell, changes, with cells becoming more elongated. In agreement with previous studies, we observe that the asymmetry in partitioning of aggregates in division decreases significantly in minimal media as well as for lower temperatures. Finally, in these sub-optimal environmental conditions, we observed a positive relationship between asymmetry in size of the daughter cell and the fraction of unwanted substances inherited in division.

We conclude that, in the sub-optimal environments tested, the larger daughter cells are also the ones that inherit more unwanted aggregates and have longer division times. Consequently, cell division in these environments usually results in a smaller but healthier daughter cell and an older, larger parent. We additionally conclude that, in sub-optimal conditions, an additional mechanism associated to cell rejuvenation becomes active. Consequently, aside from the asymmetric segregation of unwanted aggregates, the chance for morphological asymmetries in division is significantly increased, generating cells that while being smaller, are also less poised with unwanted substances

EPIGENETICS OF EARLY HUMAN T HELPER CELL DIFFERENTIATION

R. David Hawkins[1,3,7], Antti Larjo[2,5,7], Subhash K. Tripathi[3,4,7], Ulrich Wagner[6], Ying Luu[6], Tapio Lönnberg[3], Sunil K. Raghav[3], Leonard K. Lee[6], Riikka Lund[3], Harri Lähdesmäki[3,5,8], Bing Ren[6,8], and Riitta Lahesmaa[3,8]

[1]Department of Medicine, Division of Medical Genetics & Department of Genome Sciences, University of Washington School of Medicine, Seattle, Washington 98195, USA
[2]Department of Signal Processing, Tampere University of Technology, FI-33101 Tampere, Finland
[3]Turku Centre for Biotechnology, University of Turku and Åbo Akademi, FI-20520 Turku, Finland
[4]National Doctoral Programme in Informational and Structural Biology, Turku, Finland.
[5]Department of Information and Computer Science, Aalto University School of Science, FI-02150 Espoo, Finland
[6]Ludwig Institute for Cancer Research - San Diego, La Jolla, California 92093, USA & Department of Cellular and Molecular Medicine, Moores Cancer Center, and Institute of Genomic Medicine, University of California San Diego, La Jolla, California 92093, USA
[7,8]These authors contributed equally to this work
riitta.lahesmaa@btk.fi

ABSTRACT

CD4+ T cells play a crucial role in the adaptive immune system and they have the ability to differentiate to functionally distinct effector subtypes such as T helper 1 (Th1), Th2, Th17, and iTreg. Here we have studied histone modifications (H3K4me1, H3K27ac, H3K4me3) to identify the lineage-specific functional cis-regulatory elements for early differentiating human Th1 and Th2 cells. To correlate epigenetic information with gene expression, we have utilized genome-wide digital gene expression analysis from the Helicos platform. The identified enhancer regions are also overlaid with open chromatin sites (DNase-seq) from fully differentiated T cells in order to characterize whether early enhancers are active only during the early lineage specification or remain active in committed Th cells. Analysis of transcription factor binding sites at enhancers allowed us to identify known and novel transcriptional regulators, which drive the lineage determination. As improper cell fate specification can lead to immunopathogenesis, we studied the overlap of the identified enhancers with SNPs associated with different autoimmune diseases. The possibility of such overlapping SNPs to disrupt binding of TFs was studied using computational predictions and verified for a subset of cases using DAPA experiments. This study is the first looking at contribution of enhancers to early human T cell lineage specification. The obtained results also provide insight into how regulatory SNPs may contribute to disease pathogenesis.

ANALYSIS OF ALTERNATIVE SPLICING IN PROSTATE CANCER USING EXON-EXON SPLICE JUNCTIONS

Sergei Häyrynen[1], Matti Annala[2], Kati Waltering[2], Tapio Visakorpi[2], Matti Nykter[2]

[1]Institute of Signal Processing, Tampere University of Technology,
P.O. Box 553, FI-33101 Tampere, Finland
[2]Institute of Biomedical Technology, University of Tampere,
Institute of Biomedical Technology
FI-33014 University of Tampere, Finland
sergei.hayrynen@tut.fi, matti.annala@uta.fi, kati.waltering@uta.fi, tapio.visakorpi@uta.fi,
matti.nykter@uta.fi

ABSTRACT

Alternative splicing of exons may result in transcripts with varying functions. Some splicing events, such as e.g. alternatively spliced androgen receptor gene, have been shown to be associated to the emergency of the castration resistant prostate cancer (CRPC). The goal of this work was to screen RNA-sequencing data from samples of different stages of prostate cancer to find other splicing event candidates associated with the CRPC cases.

Screening based on the relative exon expressions, the ratio of reads of an individual exon and combined reads of the other exons of a gene, resulted in a list of potential splicing candidates, but especially in the samples from castration resistant prostate cancer large numbers of intronic reads introduced a level of uncertainty in the validity of relative exon expressions. To obtain more reliable and convincing results the emphasis was shifted to reads aligning to exon-exon junctions. Using Ensemble gene annotations a splice junction library of all possible exon combinations inside each gene was made. Sequenced reads from 28 untreated prostate cancer tumors, 13 CRPC cases and 12 benign prostatic hyperplasia samples were aligned with Bowtie to splice junction library in addition to continuous exonic regions in the annotations and obtained data was screened for statistically significant differing splicing events between sample groups. Proposed approach reduces the number of false positive candidates significantly.

We also compared our approach to published alternative splicing analysis methods such as MISO, a probabilistic algorithm for quantification the expression level of alternatively spliced genes (Katz et. al 2010, Nature Methods). With published tools the analysis of large number of samples proved to be problematic as many of them, such as MISO, are focused on analyzing small number of samples and comparing differences in individual samples instead of sample groups. Thus, the utilization of splice junction based approach seems more suitable to large sample cohort than a statistical model developed for two sample comparison.

SIMPLIFIED WHOLE BRAIN WHITE MATTER ANALYSIS BASED ON DIFFUSION TENSOR IMAGING

Tero Ilvesmäki[1,2], Teemu Luoto[3], Antti Brander[1], Ullamari Hakulinen[1,2], Pertti Ryymin[1], Hannu Eskola[1,2], Grant Iverson[4] and Juha Öhman[3]

[1]Medical Imaging Centre and Hospital Pharmacy, Department of Radiology, Tampere University Hospital, Tampere, Finland

[2]Department of Electronics and Communications Engineering, Tampere University of Technology, Tampere, Finland

[3]Department of Neurosciences and Rehabilitation, Tampere University Hospital, Tampere, Finland

[4]Department of Physical Medicine and Rehabilitation, Harvard Medical School; & Red Sox Foundation and Massachusetts General Hospital Home Base Program, Boston, Massachusetts, USA

tero.ilvesmaki@tut.fi, teemu.luoto@pshp.fi, antti.brander@pshp.fi, ullamari.hakulinen@pshp.fi, pertti.ryymin@pshp.fi, hannu.eskola@tut.fi, giverson@interchange.ubc.ca, juha.ohman@pshp.fi

ABSTRACT

Background: White matter (WM) abnormalities can be assessed in vivo by applying diffusion tensor imaging (DTI), a specialized imaging protocol of magnetic resonance imaging (MRI). However, methodological limitations, such as small sample sizes, failure to control for pre-injury confounding factors and a lack of analysis standards, can cause bias in the results.

Objective: To develop a procedure for quantifying WM changes connected with specific neurological diseases with enhanced reproducibility and standardization. As an application we studied whether mild traumatic brain injury (MTBI) is associated with microstructural changes in WM.

Materials and Methods: To assess the post-processing method, 75 patients with MTBI and 40 healthy control subjects were scanned under the same MRI protocol: whole-brain 3T diffusion weighted MRI, b-factor 0 and 1000 s/mm2 with 20 diffusion gradient directions. DTI parameters tested for significant differences between patient and control groups were: fractional anisotropy (FA), mean diffusivity (MD), radial diffusivity (RD) and axial diffusivity (AD). Voxel-wise statistical whole brain analysis (WBA) was carried out using tract-based spatial statistics (TBSS). The subjects' FA data were first projected onto a mean FA skeleton, representing the centers of all tracts common to the group, before applying voxel-wise intra-subject statistics. All MTBI patients and controls were compared taking age and gender into account by correlating their effects as confound regressors and by analyzing age- and gender- matched subgroups.

Results: A simplified method to operate TBSS analysis in a standardized way was developed in the form of a macro code, which merges the several TBSS scripts into one. No significant differences ($p<0.01$) were found between the control and MTBI groups or any of the subgroups for the tested DTI parameters in the application.

Conclusion: A semi-automated procedure for quantitative statistical white matter WBA was developed. Originally TBSS is split to several phases requiring user input, but the method was reduced to only one phase, which utilizes various parts of the analysis, generating complete results with only a single script. The whole brain WM analysis method was made more user-friendly, efficient and reproducible through scripting. In this study TBSS was unable to distinguish any significant white matter anomalies between groups. This can be a consequence of the statistical nature of TBSS, where localized abnormalities, characteristic to MTBI, are averaged out in the results. Without positive results the developed method was not fully utilized and quantitative results were not processed.

A DYNAMIC MODEL FOR T HELPER 17 CELL DIFFERENTIATION

Jukka Intosalmi[1], Sini Rautio[1], Helena Ahlfors[2], Zhi Jane Chen[3], Riitta Lahesmaa[3], Brigitta Stockinger[2], and Harri Lähdesmäki[1,3]

[1]Department of Information and Computer Science, Aalto University, Finland,
[2]Division of Molecular Immunology,
Medical Research Council National Institute for Medical Research, UK,
[3]Turku Centre for Biotechnology, University of Turku and Åbo Akademi, Finland,
jukka.intosalmi@aalto.fi

ABSTRACT

Introduction: The differentiation of naive CD4+ helper T cells into effector T cells is largely determined by extracellular cytokine signals. The cytokine signals activate the differentiation specific transcription factors and control the dynamics of underlying regulatory mechanisms. For T helper (Th) 17 cell differentiation, the critical cytokines are TGFβ and IL6 which activate the key transcription factors, RORγt and STAT3.

Results: In this study, we construct a mathematical model to describe the dynamics of the core components, which drive the Th17 cell differentiation. Our minimal model consists of the key transcription factors (RORγt and STAT3) and two cytokine inputs (TGFβ and IL6) driving the differentiation process. The model is implemented in the form of nonlinear ordinary differential equations to model the population average of the core components. The mRNA levels described by the model can be combined with time-course RNA sequencing data (B6 mice) through a Bayesian framework, which provides us with a data driven, probabilistic description of model parameters and outputs.

Conclusions: Our model is capable of reproducing realistic dynamics in four different cytokine conditions; for two of these conditions we have experimental time-course RNA-Seq data which our model is able to reproduce. Furthermore, our model can be used to generate predictions to hypothesize and design new wet-lab experiments. For example, we can determine cytokine conditions that lead to an increased risk of differentiation failure.

DIFFERENTIAL GENE EXPRESSION OF IMMUNOLOGICALLY ACTIVE MOLECULES BETWEEN CHILDREN BORN IN FINLAND, ESTONIA AND RUSSIAN KARELIA

Henna Kallionpää[1,2], Essi Laajala[1,5], Viveka Öling[1], Vallo Tillmann[3], Natalya V. Dorshakova[4], Harri Lähdesmäki[1,5], Mikael Knip[6,7,8], Riitta Lahesmaa[1], and the DIABIMMUNE Study Group

[1] Turku Centre for Biotechnology, University of Turku and Åbo Akademi, Turku, Finland, P.O. Box 123, BioCity, 20521 Turku,
firstname.surname@btk.fi
[2] Turku Doctoral Programme of Biomedical Sciences, Turku, Finland
[3] Department of Pediatrics, University of Tartu and Tartu University Hospital, Tartu, Estonia
[4] Petrozavodsk State University, Petrozavodsk, Russia
[5] Department of Information and Computer Science, Aalto University School of Science, Espoo, Finland
[6] Children's Hospital, University of Helsinki and Helsinki University Central Hospital
[7] Folkhälsan Research Center, Helsinki, Finland
[8] Department of Pediatrics, Tampere University Hospital, Tampere, Finland

ABSTRACT

The DIABIMMUNE project investigates environmental factors in the development of type 1 diabetes and other immune-mediated diseases in Finland, Estonia, and Russian Karelia. There are only minor differences in the frequencies of predisposing and protective HLA genotypes in these three countries, but the incidence of type 1 diabetes is six times higher in Finland compared to Russian Karelia. A wide variety of data are being collected from young children born in Espoo (Finland), Tartu (Estonia), and Petrozavodsk (Russian Karelia), including e.g. serum, RNA, breast milk samples, stool samples, and food diaries.

Our group studies the whole blood RNA samples from the DIABIMMUNE cohorts. To date, we have analyzed cord blood samples that were collected in Tempus Blood RNA tubes in Espoo (49 samples), Tartu (26 samples) and Petrozavodsk (41 samples). The RNA was isolated and GeneTitan™ instrument was used for automated hybridization on Affymetrix Human U219 array plate. The aim was to survey general differences in immunologically active molecules in these three cities. The samples had not been selected in any way, except by date of birth (1.1. − 31.5.2010), mode of delivery (children born by caesarean section were excluded) and RNA quality.

The data were pre-processed by robust multi-array average (RMA) and absent calls were filtered out by determining the threshold value empirically for each sample. Differential expression was detected by using the R Bioconductor package Limma to fit a linear model and compute a moderated t-statistic for each present probeset for all three contrasts: Espoo vs. Petrozavodsk, Tartu vs. Petrozavodsk and Espoo vs. Tartu. The results of our ongoing study suggest that some immunologically relevant differences are present between the children born in these cities. Molecular enrichment analysis revealed for example upregulated Interleukin 2 signaling in Espoo compared to Tartu.

A MULTI-PLATFORM TRANSCRIPTIONAL PROFILING PROVIDES NOVEL INSIGHTS INTO EARLY T-HELPER CELL DIFFERENTIATION

Kartiek Kanduri[1], Subhash Tripathi[1], Antti Larjo[2], Henrik Mannerström[2], Riikka Lund[1], Jane Zhi Chen[1], Harri Lähdesmäki[2] and Riitta Lahesmaa[1]

[1]Turku Centre for Biotechnology, University of Turku and Åbo Akademi University, Turku, Finland
[2]Department of Information and Computer Science, Aalto University School of Science, Helsinki, Finland
kartiek.kanduri@btk.fi, subhash.tripathi@btk.fi, antti.larjo@aalto.fi, henrik.mannerstrom@aalto.fi, riikka.lund@btk.fi, zchen@btk.fi, harri.lahdesmaki@aalto.fi, riitta.lahesmaa@btk.fi

ABSTRACT

Activation and differentiation of T-helper (Th) cells is a complex process orchestrated by distinct gene activation programs engaging a number of genes. This process is crucial for a robust immune response that even a slight imbalance might lead to disease states such as allergy or an autoimmune disease. Therefore, identification of genes involved in this processes is important to further understand the pathogenesis of immune mediated diseases. In this study we identified lineage specific genes of Th1 and Th2 subsets (at an early stage of differentiation), using both the traditional genome wide transcriptional profiling (microarrays) and next-generation sequencing techniques. Next-generation sequencing techniques do not have the limitations of the microarrays such as pre-selection bias. Results from the comparison of various transcriptomic platforms are useful for future experimental design. Importantly, these results enabled us to generate a high confidence gene list that is in agreement in all the platforms employed. We discovered also a panel of novel genes deduced from the next-generation sequencing data.

ESTOOLS DATA@HAND, A DATABASE FOR INTEGRATIVE AND COMPARATIVE STEM CELL DATA ANALYSIS

Lingjia Kong[1,2], Kaisa-Leena Aho[1*], Kirsi Granberg[1*], Riikka Lund[2*], Laura Järvenpää[1], Janne Seppälä[1], Paul Gokhale[3], Kalle Leinonen[1], Lauri Hahne[1,2], Jarno Mäkelä[1], Kirsti Laurila[1], Heidi Pukkila[1], Elisa Närvä[2], Olli Yli-Harja[1], Peter W. Andrews[3], Matti Nykter[1], Riitta Lahesmaa[2], Christophe Roos[1], Reija Autio[1,2]*

[1]Department of Signal Processing, Tampere University of Technology, Korkeakoulunkatu 1, FIN-33101, Tampere, FINLAND,
[2]Turku Centre for Biotechnology, University of Turku and Åbo Akademi University, Biocity, Tykistökatu 6, FIN-20520, Turku, FINLAND,
[3]Centre for Stem Cell Biology and the Department of Biomedical Science, University of Sheffield, Western Bank, Sheffield, S10 2TN, GREAT BRITAIN
* The authors contributed equally to this work.
lingjia.kong@tut.fi

ABSTRACT

Embryonic stem (ES) cells possess the unique and essential abilities to self-renew and to differentiate into any cell type, and have great potential in cell therapy and regenerative medicine. The amount of data generated from ES cell research increases year by year. In order to facilitate the efficient use of the data, we have developed ESTOOLS DATA@HAND (http://estools.cs.tut.fi/), a database for peer-reviewed data from microarray measurements of gene expression on ES cells. The data in ESTOOLS DATA@HAND is collected from public repositories, and contains more than 1200 samples from 77 sample sets. The data covers human embryonic stem (hES) cells, human induced pluripotent stem (hiPS) cells as well as dozens of other cell and tissue types reported in the same studies in various conditions. By integrating the published data and extensively re-annotating the samples across all experiments, the data can be combined and analyzed in new perspectives and new research questions can be asked.

All the data in ESTOOLS DATA@HAND has been preprocessed and normalized systematically across each sample set. Two sample meta-sets, Affymetrix meta-set (408 samples) and Illumina meta-set (245 samples), have been established by jointly normalized data from different studies. Thus, the integrative meta-analysis crossing experiments is possible. ESTOOLS DATA@HAND has various tools to browse, retrieve, visualize, and analyze the data, for example, to detect differentially or similarly expressed genes, to cluster the data, or to analyze enriched Gene Ontology terms and Kyoto Encyclopedia of Genes and Genomes (KEGG) pathways for a given list of genes. Further, all samples have been manually annotated along more than 60 dimensions covering biological properties and experimental parameters. This annotation information together with the various data analysis options make ESTOOLS DATA@HAND a valuable web resource for the stem cell research community.

IDENTIFYING THE ANDROGEN REGULATION NETWORK IN PROSTATE CANCER

Ville Kytölä[1], Kati Waltering[2], Tapio Visakorpi[2] and Matti Nykter[2]

[1]Department of Signal Processing, Tampere University of Technology,
P.O. Box 553, FI-33101 Tampere, Finland,
ville.kytola@tut.fi
[2]Institute of Biomedical Technology, University of Tampere,
Kalevantie 4, 33014, Tampere, Finland

ABSTRACT

Prostate cancer is one of the most common cancers in the modern world. One of the key identified elements in the development and progression of the disease is the function of the androgen receptor (AR). Despite the extensive research concerning the regulatory role of the AR, the regulation cascade induced by the androgen receptor has not been thoroughly characterized.

Here we use computational analysis of time series microarray data to uncover the temporal regulation of AR using data from Massie et al (EMBO J. 2011). First, we identified transcription factors (TF) whose expression change significantly during the first hours after hormone stimulation in LNCaP cells. Subsequently, the direct targets of these TFs were identified by binding prediction and literature curation. K-means clustering was performed for the expression dataset to identify modules of co-expressed genes.

Two independent methods were used to link the early reacting transcription factors to clusters they putatively regulate. First, a gene set enrichment analysis was used for the targets of each early reacting TF in each of the clusters. Secondly, a time-lagged correlation based method was applied to the clusters in order to find correlation connections with reacting TFs.

As a result, a cascade of regulatory interactions between AR primary and secondary targets was obtained. The early reacting genes, which presumably were direct AR targets seemed to regulate specific clusters of genes. These secondary targets can be used for further explore the regulatory and signaling pathways that propagate the AR signaling. Experimental validation of the predicted cascades is ongoing

COMBINATORIAL REGULATION OF LIPOPROTEIN LIPASE BY MICRORNAS DURING MOUSE ADIPOGENESIS

Maria Liivrand[1,2], *Merja Heinäniemi*[1], *Elisabeth John*[1], *Jochen G. Schneider*[2,3], *Thomas Sauter*[1#] *and Lasse Sinkkonen*[1#*]

[1]Life Sciences Research Unit, University of Luxembourg, L-1511 Luxembourg, Luxembourg

[2]Luxembourg Centre for Systems Biomedicine, University of Luxembourg, L-4362 Esch-sur-Alzette, Luxembourg

[3] Saarland University Medical Center, Dpt. of Medicine II, Homburg, Saar, Germany

[#] The authors contributed equally

[*] Corresponding author: lasse.sinkkonen@uni.lu

ABSTRACT

MicroRNAs (miRNAs) regulate gene expression directly through base pairing to their targets or indirectly through participating in multi-scale regulatory networks. Often miRNAs take part in feed-forward motifs where a miRNA and a transcription factor act on shared targets to achieve accurate regulation of processes such as cell differentiation. Here we show that the expression levels of miR-27a and miR-29a inversely correlate with the mRNA levels of lipoprotein lipase (*Lpl*), their predicted combinatorial target, and its key transcriptional regulator peroxisome proliferator activated receptor gamma (*Pparg*) during 3T3-L1 adipocyte differentiation. More importantly, we show that *Lpl*, a key lipogenic enzyme, can be negatively regulated by the two miRNA families in a combinatorial fashion on the mRNA and functional level in maturing adipocytes. This regulation is due to direct targeting of the Lpl 3'UTR as confirmed by reporter gene assays. In addition, a small mathematical model captures the dynamics of this feed-forward motif and predicts the changes in *Lpl* mRNA levels upon network perturbations. The obtained results might offer an explanation to the dysregulation of LPL in diabetic conditions and could be extended to quantitative modeling of regulation of other metabolic genes under similar regulatory network motifs. To obtain a global overview of dynamic expression profiles of differentiation in adipocytes and related lineages, integration of multiple genome-wide approaches will be further applied.

RE-PLOT: WEB APPLICATION FOR GENOMIC DATA ILLUSTRATION

Jake Lin[1,2,3], Richard Kreisberg[1], Aleksi Kallio[2], Patrick May[1,3], Olli Yli-Harja[2], Matti Nykter[4], Ilya Shmulevich[1], Reija Autio[2]

[1] Institute of Systems Biology, Seattle, USA
[2] Department of Signal Processing, Tampere University of Technology, Finland
[3] Luxembourg Centre for Systems Biomedicine, University of Luxembourg, Luxembourg
[4] Institute of Biomedical Technology, University of Tampere, Finland
jake.lin@systemsbiology.org, reija.autio@tut.fi

ABSTRACT

Systems biology experiments are studying wide range of topics with various organisms often using next generation technologies to produce thousands of data points across different -omics data types, including RNA, methylation, protein, gene, metabolite and copy number metrics. When analyzing this kind of data, the data mining algorithms and statistical analyses are yielding ranked and heterogeneous results and association networks distributed over the entire genome. To assist with the exploration and visualization of these results, we have developed RE-Plot, a web application to integrate and visualize genomic feature relationships and annotations across a broad range of organisms - human, mouse, nematode, fly, yeast, zebra fish, Arabidopsis and rice.

Custom chromosomal feature networks can be uploaded as simple text files into the RE-Plot, and an optional numeric score column, such as p-value or correlation for filtering, is supported. In addition, RE-Plot supports all pairing of genomic types and the end nodes are color-coded based on its type. The graph represents the chromosome lengths as perimeter segments, as a reference outer ring, such as cytoband for human. The inner arcs between nodes represent the uploaded network allowing also filtering if the optional score column is included. Multiple annotation rings such as depiction of highly mutated versus expressed regions, can be uploaded as text files. These genomic annotations can be associated with values and visualized as continuous histogram rings. RE-Plot interacts with genomic browsers and clicking on applicable nodes will launch the UCSC Genome Browser for human and mouse and gbrowse for zebra fish and Arabidopsis. Other supported visualization views are tabular and cytoscapeweb network layouts. The visualization outputs can be exported as svg image or tabular text files. The freely available RE-Plot is built using open sourced JavaScript, HTML5, python and SQLite.

UNCOVERING DEVELOPMENTAL LINEAGES OF HEMATOLOGICAL MALIGNANCIES

Thomas Liuksiala[1], Merja Heinäniemi[2], Kirsi Granberg[1], Matti Nykter[1] and Olli Lohi[3]

[1]Department of Signal Processing, Tampere University of Technology,
[2]A. I. Virtanen Institute for Molecular Sciences, University of Eastern Finland
[3]Paediatric Research Centre, University of Tampere
Department of Signal Processing, Tampere University of Technology,
P.O. Box 553, FI-33101 Tampere, Finland,
thomas.liuksiala@tut.fi

ABSTRACT

Blood cells are produced by hematopoiesis, in which hematopoietic stem cells give rise to mature cell types, each having its own task in the circulatory or immune system. When a stem cell, precursor cell or fully differentiated blood cell becomes cancerous, the resulting condition is called myeloma, leukemia, or lymphoma. The classification scheme of blood cancers, or hematological malignancies, has conventionally relied on clinical traits. Recently, more emphasis has been given to the lineage of the cell type from which the cancer originates.

Using microarray gene expression data from 6000 cancer and 900 normal blood cell samples we computationally linked blood cancer subtypes to normal cell types. The underlying hypothesis was that the normal cell type with the most similar gene expression profile to a cancer subtype is the most likely origin of the malignant behavior. The analysis was carried out with two different gene subsets: genes with the highest variance over all samples and transcription factors. Principal component analysis was used to reduce data dimensionality. Each cancer sample was linked to a normal cell type by k-nearest neighbor classification. The percentage of samples of a specific cancer type classified to a normal cell type is interpreted as the similarity of the cancer and the normal cell.

The results, so far, reflect fairly well the known biology of hematological malignancies. Myeloid leukemias, for instance, were clearly identified as myeloid originating and lymphomas were strongly linked to lymphocytes. Somewhat surprisingly, though, lymphomas were also often linked to macrophages. Rather than cell of origin, this suggests high level of macrophage activation in lymph node tumor samples. In addition, many leukemias showed a signature of hematopoietic stem cells, indicating a transition towards stem cell like state as the result of cancer development.

EFFECT OF ENVIRONMENTAL STRESS ON THE *IN VIVO* KINETICS OF SEGREGATION OF UNWANTED PROTEIN AGGREGATES IN *E. COLI*, AT SINGLE CELL, SINGLE EVENT LEVEL.

Ramakanth Neeli Venkata[1], *Abhishekh Gupta*[1], *Anantha-Barathi Muthukrishnan*[1], *Olli-Yli Harja*[1, 2] *and Andre S Ribeiro*[1,*]

[1]Laboratory of Biosystem Dynamics, Computational Systems Biology Research Group
Department of Signal Processing, Tampere University of Technology,
P.O. Box 553, FI-33101 Tampere, Finland,

[2]Institute for Systems Biology, 1441N 34th St, Seattle, WA, 98103-8904, USA.

ramakanth.neelivenkata@tut.fi

ABSTRACT

Recent studies in *Escherichia coli* reported that these organisms are capable of, when dividing, biased partitioning of unwanted substances by the daughter cells, as a means to cope with aging. However, little is known about the possible effects of stressful environmental conditions on this process, namely, on the kinetics of segregation of the unwanted protein aggregates to the cell poles and subsequent partitioning in division. We use an RNA-MS2-GFP tagging method to study the *in vivo* partitioning in division of unwanted aggregates. We acquired fluorescent confocal time lapse images as cells divide and accumulate these aggregates, for several hours. By employing a microfluidics system, we subject cells to various stress conditions such as nutrient deprivation, temperature shift, oxidative stress, and hyperosmotic shift. Also, we performed qPCR as a means to assess the stress response levels in the conditions tested by using the *rpoS* as a target gene. From the data, we report our findings on the bias and kinetics of partitioning of unwanted substances as a function of the stress levels. The results should provide better understanding of how these organisms cope with stress.

THE TUMORIGENIC FGFR3-TACC3 GENE FUSION ESCAPES MIR-99A REGULATION IN GLIOBLASTOMA

Brittany Parker [1,2], Matti Annala [1,3,4], David Cogdell [1] , Kirsi Johanna Granberg [1,3,4], Yan Sun [1,5], Ping Ji [1],Xia Li [1], Joy Gumin [1], Hong Zheng[5], Limei Hu [1], Olli Yli-Herja[3], Hannu Haapasalo [4,6], Tapio Visakorpi [4], Xiuping Liu[1], Chang-gong Liu[1], Raymond Sawaya[1], Gregory Fuller[1], Kexcin Chen[5], Fredrick Lang[1,2], Matti Nykter [3,4] and Zhang Wei [1,3]

[1]The University of Texas MD Anderson Cancer Center, 1515 Holcombe Blvd Houston, TX 77030
[2]The University of Texas Graduate School of Biomedical Sciences, Houston, Texas, USA
[3]Tampere University of Technology, Tampere, Finland
[4]University of Tampere, Tampere, Finland
[5]Tianjin Medical University Cancer Institute and Hospital, Tianjin, People's Republic of China
[6]Tampere University Hospital, Tampere, Finland
brittanycparker@gmail.com

ABSTRACT

Fusion genes are chromosomal aberrations that are found in many cancers and can be used as prognostic markers and drug targets in clinical practice. Fusions can lead to production of oncogenic fusion proteins or to enhanced expression of oncogenes. Several recent studies have reported that some fusion genes can escape microRNA regulation via 3'–untranslated region (3'-UTR) deletion. We performed whole transcriptome sequencing to identify fusion genes in glioma and discovered FGFR3-TACC3 fusions in 4 of 48 glioblastoma samples from patients both of mixed European and of Asian descent, but not in any of 43 low-grade glioma samples tested. The fusion, caused by tandem duplication on 4p16.3, led to the loss of the 3'-UTR of FGFR3, blocking gene regulation of miR-99a and enhancing expression of the fusion gene. The fusion gene was mutually exclusive with EGFR, PDGFR, or MET amplification. Using cultured glioblastoma cells and a mouse xenocraft model, we found that fusion protein expression promoted cell proliferation and tumor progression, while WTFGFR3 protein was not tumorigenic, even under forced overexpression. These results demonstrated that the FGFR3-TACC3 gene fusion is expressed in human cancer and generates an oncogenic protein that promotes tumorigenesis in glioblastoma.

GENE LOCATION AND PROXIMITY IN BACTERIAL GENE REGULATION

Otto Pulkkinen[1] *and Ralf Metzler*[1,2]

[1]Department of Physics, Tampere University of Technology, Finland
[2]Institute for Physics & Astronomy, University of Potsdam, Germany
otto.pulkkinen@tut.fi

ABSTRACT

The tendency of bacterial transcription factor (TF) genes and their binding sites to colocalize has been known for a long time. One hypothesis for the observed colocalization is based on possibly shorter search times for proximal binding sites in comparison to distal ones. On the one hand, diffusion of signaling molecules in the cytosol has been observed to be fast, so one might assume that the spatial distributions of TFs is uniform and that these aspects can be neglected, but on the other hand, in a recent study by Kuhlman and Cox [Mol. Syst. Biol. 8, 610 (2012)], considerable spatial variation in intracellular TF concentrations was observed, and its effect on regulation recognized.

We present a general, analytical theory for bacterial gene regulation in the case that the TF is a repressor. We consider several aspects of the regulation process: transcriptional and translational stochasticity in TF production, the transport of TF molecules to their binding site by facilitated diffusion in a spatially structured cellular environment, and the nonspecific binding of TFs to the DNA near the binding site. We provide analytical formulae for the mean and variance of the target gene transcription rate.

We show from analytical and numerical analysis that the distance between a transcription factor gene and its target gene can drastically affect the speed and reliability of transcriptional regulation. The observed variations in regulation efficiency are linked to the magnitude of the variation of the TF concentration peaks as a function of the binding site distance from the signal source. Finally, we discuss transcriptional pulsing and the effect of TF gene location on the transcription rate of the target gene.

IDENTIFICATION OF CANCER TYPES WITH NRF2 HYPERACTIVITY

Petri Pölönen[1], Antti Ylipää[2], Matti Nykter[3], Anna-Liisa Levonen[1] and Merja Heinäniemi[1]

[1]Department of Biotechnology and Molecular Medicine, A.I. Virtanen Institute for Molecular Sciences, University of Eastern Finland, P.O. Box 1627, FIN-70211, Kuopio, Finland
[2]Department of Signal Processing, Tampere University of Technology, Korkeakoulunkatu 1, 33200 Tampere, Finland
[3]Institute for Biomedical Technology, University of Tampere, Biokatu 8, 33520 Tampere, Finland
petri.polonen@uef.fi

ABSTRACT

Nuclear factor (erythroid-derived 2)-like 2 (NRF2) is a transcription factor, which senses oxidative and electrophile stress. When activated, NRF2 accumulates in the nucleus where it induces the expression of cytoprotective target genes. Accumulating evidence suggests that constitutively active NRF2 has a pivotal role in cancer as it induces pro-survival genes that promote chemoresistance and cancer cell proliferation. Therefore NRF2 is a novel oncogenic transcription factor, but the prevalence on NRF2 dysregulation and functions in cancer have not been fully characterized. We analyzed microarray data of over 900 cancer cell lines in Cancer Cell Line Encyclopedia (CCLE) and created a NRF2 signature model based on our previous microarray data to identify cancers with overactive NRF2 status. Four novel cancer types and a total of 78 cancer cell lines were discovered by two individual tools to have overactive NRF2 with > 95 % probability or FDR of 0.01. Furthermore, we investigated cancer types with constitutively active NRF2 in clinical samples available in The Cancer Genome Atlas (TCGA) and found characteristic NRF2 signature also in multiple TCGA samples. We also studied the TCGA mutation data to assess whether specific mutations could explain the mechanism of NRF2 hyperactivity. TCGA data can also reveal altered gene copy numbers and miRNA profiles, which may relate to the potential mechanism of increased NRF2 activity.

SCREENING THE PROSTATE CANCER SUSCEPTIBILITY LOCI AT 2Q37.3 AND 17Q12-Q21 FOR NOVEL CANDIDATE GENES IN FINNISH PROSTATE CANCER FAMILIES

Tommi Tapani Rantapero[1], Virpi Laitinen[1], Daniel Fischer[2], Elisa Vuorinen[1], Tiina Wahlfors[1] and Johanna Schleutker[3]

[1]Institute of Biomedical Technology, FI-33014 University of Tampere, Finland

[2]School of Health Sciences, University of Tampere, Finland

[3]Medical Biochemistry and Genetics, Institute of Biomedicine, University of Turku, Finland

tommi.rantapero@uta.fi

ABSTRACT

According to several studies, genetic risk factors have been shown to be associated to prostate cancer susceptibility. Several chromosomal loci have been shown to be associated to familial prostate cancer. In a recent genome-wide linkage study strong signals coming from 2q37 and 17q21-22 were discovered in Finnish population. To study these loci in detail we performed a targeted high-throughput DNA sequencing on 21 families including 65 cases and 5 controls. In addition, RNA-sequencing was performed for 33 of these cases from purified RNA from whole blood. The aim of this study was to identify variants that associate to prostate cancer susceptibility.

Variant calling from sequencing was done using Samtools and variants were subsequently annotated using information from UCSC genome browser database. Three pathogenicity prediction tools Polyphen-2, Pon-P and Mutation taster were used to elucidate the possible phenotypic effects of variants located within genes. As an alternative approach to prioritize variants for validation in a larger population, a search for possible prostate cancer associated genes within the regions of interest was done utilizing information gathered from literature, Gene-Ontology and pathway databases, and Cancer Gene databases. To study the intergenic variants in more detail an eQTL-analysis was conducted applying two statistical models: Linear and a non-parametric directional test based model.

As a result of the pathogenicity prediction 152 variants, which might affect protein function, were discovered. From this set of variants 38 in addition with 20 additional variants from selected candidate genes were chosen for further validation which is currently ongoing.

CLASSIFICATION AND ERROR ESTIMATION FOR INDIRECT IMMUNOFLUORESCENCE IMAGES OF HEP-2 CELLS

Pekka Ruusuvuori, Tapio Manninen and Heikki Huttunen

Department of Signal Processing, Tampere University of Technology,
P.O. Box 553, FI-33101 Tampere, Finland,
pekka.ruusuvuori@tut.fi, tapio.manninen@tut.fi, heikki.huttunen@tut.fi

ABSTRACT

Indirect immunofluorescence (IIF) imaging can be used for studying protein localization in diagnosis of autoimmune diseases. In the analysis of IIF microscopy images the key challenge is the interpretation of fluorescence patterns, corresponding to protein localizations in cells. The patterns can be used for cell classification and for aiding diagnosis, but interpretation of fluorescence patterns is challenging as the image quality varies, and thus, even expert labeling is not fully reliable. Automated methods for classifying the cells are needed in order to enable routine use of computer aided diagnostics to support subjective analysis and in order to make the pattern interpretation more consistent.

Lately, efforts for increasing the interest in classifying cells in IIF images through organizing contests and sharing datasets, particularly in the context of HEp-2 cells classification, have led to emergence of several classification methods. Typically these methods extract a set of features from the original IIF images and use statistical pattern recognition tools for predicting the class, in other words, cell type for an individual object in image. Building such classification methods is a two-fold challenge: First, the classifier should be able to deal with noisy, low contrast image data with highly variable intensity levels and pattern distributions where some of the cell types are difficult to distinguish. Second, the amount of data available for training is limited since obtaining expert labeling is costly. The first challenge leads to the use of complex classifiers which are known to be prone to overfitting, while the second challenge of having limited datasets makes it difficult to generate algorithms that would generalize well. Here we describe an approach for classifying HEp-2 cell types from IIF images while keeping in mind the generalizability through presenting several error estimates.

The cell classification approach involves a feature extraction step, where hundreds of features (related to, e.g., intensity, texture, shape, local binary patterns, histograms of oriented gradients) are quantified and normalized before feeding them to the logistic regression and support vector machine classifiers. The data, originally from the ICPR 2012 Contest on HEp-2 Cells Classification (http://mivia.unisa.it/hep2contest/), was divided into training and test sets, and the classifier function was trained in nested cross validation loop where classifier parameters were optimized before doing the actual cross validation error estimation. We used 10-fold CV and leave-one-out both in cell and image level for error estimation. Furthermore, we examined the correspondence between the obtained error estimates and the error levels obtained for test data.

STUDY OF *IN VIVO* TRANSCRIPTION DYNAMICS OF *LAC* PROMOTER AT SINGLE CELL LEVEL

Adrien Sala, Stefania Garasto, Meenakshisundaram Kandhavelu and Andre Sanches Ribeiro

Laboratory of Biosystems Dynamics, Tampere University of Technology,
Department of Signal Processing, Tampere University of Technology,
P.O. Box 553, FI-33101 Tampere, Finland,
adrien.sala@tut.fi

ABSTRACT

Gene expression is a complex and stochastic process, which is regulated at different levels of transcription and translation. Previous studies suggest that gene expression in prokaryotes is controlled mainly at the level of transcription initiation. Here, we study the dynamics of mRNA production under the control of P_{lac} promoter at the single cell level using time lapse microscopy and signal processing methods. We use the MS2-GFP (RNA-protein tagging) method to detect the production of mRNA. In this method, a single copy BAC vector contains P_{lac} tagged with 48 binding sites for the MS2d coat protein. Another plasmid, which encodes the production of MS2d-GFP under the control of $P_{LtetO-1}$, was used as a reporter. The transcriptional activity of P_{lac} can be observed as fluorescent spots inside the cells. Cells containing the plasmids were treated with IPTG and aTc to induce P_{lac} and $P_{LtetO-1}$ respectively. Fluorescent images of the cells were acquired using a confocal microscope to detect the spots in the cells. From the images, using image analysis, we extracted the number of RNA molecules produced, mean and standard deviation of number of RNA molecules per cell, interval between subsequent transcription events, and number and duration of steps in transcription. From the data, an analysis of the results is provided, concerning the noise of this process. Also, we compare the activity of this promoter with that of other promoters, recently studied using the same methodology.

NOVELLETTE: A PIPELINE FOR NOVEL TRANSCRIPT AND GENE STRUCTURE IDENTIFICATION FROM RNA-SEQ DATA

Janne Seppälä[1,2], Matti Annala[1,2] and Matti Nykter[2]

[1]Department of Signal Processing, Tampere University of Technology, Tampere, Finland
[2]Institute of Biomedical Technology, University of Tampere, Tampere, Finland
janne.seppala@tut.fi

ABSTRACT

RNA-sequencing has become a standard method for quantifying mRNA expression accurately in cancer and disease studies. A typical RNA-seq pipeline consists of a read aligner and a number of custom tools and scripts to extract expression values for both known genes and novel transcripts from aligned read counts. In addition, several distinct tools to determine typical gene structure features (ORFs, TATA box, poly-A tail motif etc.) from DNA or RNA sequences of the identified transcripts are often used. However, a single, easy-to-use and flexible tool that utilizes both alignment data and gene structure identification for quantitative and qualitative assessment of novel transcripts has not yet been published.

In this work, Novellette – a tool for thorough novel transcript analysis – is presented. In short, Novellette first extracts interesting regions in the genome (e.g. differentially expressed regions between cancerous and healthy samples), then filters out regions that overlap with known genes, and finally performs a gene structure analysis for the remaining regions, resulting in a list of novel transcript candidates with their gene structure features. To test Novellette, an RNA-seq analysis of both novel transcripts and a subset of known genes is performed with publicly available data. The results show that Novellette is able to correctly reconstruct a known gene and identify the typical structural features of protein coding genes, when only a single exon of the gene is given as input. Furthermore, Novellette reliably estimates the significance of identified novel transcripts based on their structural features and outputs a sorted list of novel transcript candidates.

A FINITE ELEMENT METHOD FOR THE SIMULATION OF LAMELLIPODIUM DYNAMICS

Nikolaos Sfakianakis[1]

[1] Department of Mathematics, University of Mainz, Germany

ABSTRACT

The cytoskeleton is a cellular skeleton inside the cytoplasm of living cells. The front of the cytoskeleton, also known as lamellipodium, is the driving mechanism of the cells motility [1]. The lamellipodium is comprised by long double helix chains of actin protein termed actin-filaments.

The filaments have one of their ends at the cell membrane (plus end), and their other end (minus end) inside the cytoplasm. They polymerize by addition of actin monomers in their plus end and at the same time they depolymerize at their minus end. They are inextensible, behave like elastic beams with friction to the substrate. They exhibit crosslinks between each other and are subjects of forces exerted on them by the membrane as well as contractile forces caused by myosin. Moreover, new filaments are nucleated at the membrane.

We develop in this work a Finite Element method for the simulation of the both stationary -not moving but still highly dynamic structures- as well as moving lamellipodia.

The model we resolve numerically was proposed in [2] and [3] and is 4th order parabolic delay problem. It assumes that the lammelipodium is comprised of two families of 2 dimensional filaments. The *System* that stems reads as:

$$\underbrace{\mu^B \partial_s^2 \left(\eta^{\pm} \partial_s^2 F^{\pm}\right)}_{\text{bending}} - \underbrace{\partial_s \left(\eta^{\pm} \lambda^{\pm} \partial_s F^{\pm}\right)}_{\text{in-extensibility}} + \underbrace{\eta^{\pm} \mu^A D_t^{\pm} F^{\pm}}_{\text{adhesion}} \pm \underbrace{\partial_s \left(\eta^{+}\eta^{-}\mu_{\pm}^T(\varphi - \varphi_0)\partial_s F^{\pm^{\perp}}\right)}_{\text{twisting}}$$
$$\pm \underbrace{\eta^{+}\eta^{-}\mu^S \left(D_t^{+}F^{+} - D_t^{-}F^{-}\right)}_{\text{stretching}} = 0$$

where with $\pm$ we denote the two families of filaments, $F^{\pm}(s, \alpha, t) \in \mathbb{R}^2$, $s \in [0, L]$, $\alpha \in [0, 2\pi]$, $t > 0$ describes the position of the filament α of the family $\pm$ at time t, L is the maximal length of the filaments, $\eta^{\pm}$ are distributions functions that for the graded length, ϕ_0 is the preferred angle of the crosslinked filaments, ϕ their actual angle, and μ^B, μ^A, μ^T, μ^S are the state parameters of the problem. $D_t = \partial_t - v\partial_s$ is the material derivative operator where v is the polymerization speed.

The discretization of the *System* is with respect to (s, a) and $t > 0$.We have used two dimensional finite element method with Hermite basis function along the s-direction, and Lagrange basis along the a-direction.

The non-linearity in the in-extensibility term is treated by an implicit-explicit discretization; this gives rise to two more equations for $\lambda\pm$. The adhesion term is discretized explicitly in time. The stretching and twisting terms couple the two families; the temporal derivatives in the stretching term are treated by a predictor-corrector step.

We refer to the links, below for two numerical tests. In the first we consider a rotational symmetric lamellipodium, and the second one is the twisting term has been deactivated. In both examples, the computational domain has been discretized with 32 filaments in the α direction and 9 nodes in the s direction. A tangential and inner-directed force has been implemented as inner boundary condition, modeling myosin pulling forces.

- http://homepage.univie.ac.at/nikolaos.sfakianakis/files/LamEquilibrium.mp4
- http://homepage.univie.ac.at/nikolaos.sfakianakis/files/LamZeroTwisting.mp4

References

[1] J. Small, T. Stradal, E. Vignal, and K. Rottner, "The lamellipodium: where motility begins," *Trends Cell Biol.*, vol. 12, pp. 112–20, 2002.

[2] D. Oelz, C. Schmeiser, and J. Small, "Modelling of the actin-cytoskeleton in symmetric lamellipodial fragments," *Cell Adhesion and Migration*, pp. 117–126, 2008.

[3] D. Oelz and C. Schmeiser, *How do cells move? mathematical modelling of cytoskeleton dynamics and cell migration*. Chapman and Hall / CRC press 2009, 2009.

SENSITIVITY ANALYSIS OF SIGNALING PATHWAYS - STANDARD METHODS, NONSTANDARD INTERPRETATION

Jaroslaw Smieja

Institute of Automatic Control, Silesian University of Technology,
Akademicka 16, 44-100 Gliwice, Poland,
Jarosław.Smieja@polsl.pl

ABSTRACT

Following rapid increase of biological data in recent years, mathematical modeling of signaling pathways evolved into a large area of research. Many models have been developed so far, describing different pathways, ranging from very simple to very detailed, high order systems.

There are many different methods that can be used to describe such systems and their choice is subject to a particular question that the analysis should answer. In this work, ordinary differential equations that describe concentration of molecules involved in the pathway will be used. This gives a rise to a high dimensional model with a large number of parameters, that are unknown and difficult to estimate. Therefore, each model should be checked with respect to its sensitivity to parameter changes. Hence, sensitivity analysis became one of the necessary tools in investigation of signaling pathways, as it provides information not only about dependence between parameter values and system behavior, but also about robustness of these systems.

While sensitivity methods have been successfully applied to analysis of various pathways, they dealt with simulation results whose units were clearly determined (as concentration units). It is a reasonable approach if one wants to evaluate, for example, variation of cellular responses due to heterogeneity of cell population or the model is built on fully quantitative data coming from experiments providing absolute values of measured quantities. Unfortunately, the latter is not the case in molecular biology. In most cases available data, though quantitative, is relative (i.e. available information is about the fold increase of the number or concentration of given molecules and not about their absolute values). Therefore, to allow for comparison of experimental data coming from different sources and simulation results, normalization of the results, both experimental and numerical, is necessary. Implications of such normalization, as far as sensitivity analysis is concerned, are not discussed in the literature.

This work summarizes known sensitivity indices with regard to their applicability in analysis of models built on normalized experimental data. Additionally, it introduces another step to sensitivity analysis, based on frequency distribution of the system output. Thus, it makes possible analysis of sensitivity in pathways whose elements oscillate with various frequencies (e.g. when two oscillating regulatory modules are combined into one model, or dynamics of a pathway is analyzed in the context of a cell cycle or a circadian clock).

This work was partially supported by the NCN grant DEC-2012/04/A/ST7/00353”

UNCOVERING THE ASSOCIATIONS BETWEEN THE HOST GENOTYPE AND THE GUT MICROBIOTA

Juhi Somani[1], Noora Alakulppi[2], Jarno Tuimala[2], Timo Erkkilä[4], Päivi Saavalainen[3], Pirjo Wacklin[2], Harri Lähdesmäki[1]

[1]Department of Information and Computer Science, Aalto University, PO Box 15400, FI-02150 Espoo, Finland,
[2]Finnish Red Cross Blood Service, FI-00310 Helsinki, Finland,
[3]Research Programs Unit, Immunobiology, University of Helsinki, PO 21, Helsinki, Finland,
[4]Amazon.com, Seattle, WA 98109, USA,
juhi.somani@aalto.fi, Noora.Alakulppi@veripalvelu.fi, jtuimala@gmail.com, erkkila@amazon.com, paivi.saavalainen@helsinki.fi, Pirjo.Wacklin@veripalvelu.fi, harri.lahdesmaki@aalto.fi

ABSTRACT

The human gut microbiota is highly variable from person to person, but many studies have been conducted to examine as to what extent host genetics control the composition. Candidate gene approaches, in which one gene is deleted or added to a model organism, have been successful to show that a single host gene can have a tremendous effect on the diversity and population structure of the gut microbiome. In contrast to the candidate gene approach, the aim of this study is to assess these genotypic associations on a large-scale in human.

For 71 healthy Finnish individuals, the host genomics (from blood derived DNA) was analyzed using the Illumina ImmunoChip SNP genotyping platform. The bacterial composition of the gut (from faecal samples) was extracted applying bar-coded pyrosequencing to the V1-V3 region of 16s RNA genes, where the sequences were further binned into operational taxonomic units (OTUs). To find associations between the host genotype and their corresponding bacterial composition, we employed various statistical and computational techniques. We opted for random forests, pair-wise linear modeling and one-way analysis of variance (ANOVA). Furthermore, several dimension reduction methods such as principal component analysis (PCA), diversity indices and haplotypic blocking, were adopted to reduce dependencies and noise within both the genotype as well as bacterial data.

By applying the diverse set of tools, a number of SNPs from host genotype were found to be at least weakly associated to the gut microbiota. We continued by mapping the detected SNPs to their closest genes and then carried out pathway and ontology enrichment analysis by adjusting the background gene set according to the design of the ImmunoChip. As a result, the detected pathways and ontologies, which were either strongly or weakly enriched include, among others, immune response, Crohn's disease and systemic lupus erythematosus. Moreover, it was noticed that a handful of bacteria might be responsible for approximately 95% of the variation in the bacterial abundances across samples. This may give an idea as to which bacteria's abundances are targeted genotypically. Ergo, by using various statistical and computational techniques, a better understanding of how the gut microbiota are assembled, maintained and associated to the host genotype can be acquired.

This study was supported by the SalWe Research Program for IMO (Tekes - the Finnish Funding Agency for Technology and Innovation grant 648/10). Moreover, the SNP annotation came from T1DBase in collaboration with Dr. Oliver Burren.

PRIMARY miRNA ANNOTATION FROM GRO-SEQ DATA

Liisa-Ida Sorsa[1,2], Merja Heinäniemi [3] , Matti Nykter[2]

[1]Department of Signal Processing, Tampere University of Technology,
P.O. Box 553, FI-33101 Tampere, Finland
[2] Institute of Biomedical Technology, University of Tampere, 33520 Tampere, Finland
[3] A.I. Virtanen Institute for Molecular Sciences, University of Eastern Finland, 70211, Kuopio, Finland
liisa-ida.sorsa@tut.fi

ABSTRACT

MicroRNAs (miRNAs) play important roles in gene regulation networks. These molecules are transcribed by RNA polymerases analogously to protein coding genes. However, many miRNA are processed co-transcriptionally making the annotation of primary transcripts, which can be tens of kilobases in length, problematic. Accurate annotation of primary transcripts is necessary for example to study transcriptional regulation of miRNA. Current annotation efforts are based on infering the putative transcriptions start sites (TSS) based on various histone modifications from ChIP-seq experiments. Here we use recently developed Global run-on sequencing (GRO-seq), which maps the position, amount and orientation of transcriptionally engaged RNA polymerases genome-wide, to study nascent RNA with goal to identify primary miRNA transcripts.

To find TSS from GRO-seq data, we have developed a data-driven method for processing GRO-seq signal. We apply our method to HUVEC cells to uncover primary sequence for miRNA that are transcribed in these cells. We calculated the number of coding strand reads at each genomic position to identify the primary transcripts. We then determined the near-symmetrical, divergent peaks to identify the TSSs. The primary transcript is interpreted as the non-zero signal starting from the TSS. Comparison with annotated protein coding transcripts and histone modification data from ENCODE shows that TSS and transcript bodies can successfully be uncovered from GRO-seq data, enabling accurate identification of primary miRNA transcripts.

A LOWER BOUND FOR THE CONFIDENCE INTERVAL OF THE MUTUAL INFORMATION OF HIGH DIMENSIONAL RANDOM VARIABLES

Arno G. Stefani[1], Johannes B. Huber[1], Christophe Jardin[2] and Heinrich Sticht[2]

[1]Institute for Information Transmission (LIT), FAU Erlangen-Nuremberg,
Cauerstr. 7, 91058 Erlangen, Germany
[2]Bioinformatics, Institute for Biochemistry, FAU Erlangen-Nuremberg,
Fahrstr. 17, 91054 Erlangen, Germany
{stefani, huber}@LNT.de, {christophe.jardin, h.sticht}@biochem.uni-erlangen.de

ABSTRACT

Introduction: Given an i.i.d. sample of pairs (x_i , y_i), i = 1, . . . , n, of two random variables X, Y, the mutual information (MI) I(X; Y) (see [1]) is often hard to estimate if one or both of the two random variables are high dimensional and nothing is known about their joint distribution, particularly for small sample sizes. E.g. for a sample size n = 1000, X 10-dimensional and binary, Y 1-dimensional and binary, Y would already be divided into $2^{10} = 1024$ partitions and the space of X and Y together into 2048 partitions, and therefore any estimated confidence interval would be quite large due to the small sample size compared to the number of partitions. In this paper this problem is solved by using the k-means (see [2]) algorithm to get a specified number of partitions for X and Y and afterwards applying a suitable confidence interval estimator.

Results: In this paper it is assumed that $X = (X_1 , X_2 , . . . , X_{nx})$ and $Y = (Y_1 , Y_2 , . . . , Y_{ny})$ are multivariate random variables, where each component can be discrete or continuous. Then the following observation is made: As soon as at least one component of X and Y is continuous or countably infinite, and nothing is known about the joint distribution of (X, Y), the MI I(X; Y) can be anything from 0 to ∞. Therefore, in this situation, it is impossible to find an upper bound of the MI confidence interval, and the best one can hope for is a lower bound.

This lower bound can be calculated by the following steps. First the k-means algorithm is applied independently on the i.i.d. sample of X and Y, (x_i), (y_i), i = 1, . . . , n, yielding the new random variables X^q and Y^q (q for quantized) which are 1-dimensional, discrete and have freely chosen alphabet sizes k_x and k_y. By the data processing theorem (see [1]) $I(X; Y) \geq I(X^q ; Y^q)$. Next one of the confidence interval estimators described in [3] and [4] is used to find the desired lower bound on the MI confidence interval.

Now the question remains how to choose k_x and k_y. Here simply the right tradeoff between bias and variance has to be found. Small k_x and k_y results in a large loss of information during the application of the k-means algorithm. Large k_x and k_y result in a large variance and therefore a wide confidence interval. Somewhere in between is an optimal pair k_x , k_y that gives the best lower bound, and this lower bound can be found by varying k_x and k_y, starting with $k_x = 2$, $k_y = 2$ and calculating the lower bound for all pairs k_x and k_y until the lower bound starts to drop significantly.

The authors have applied the described for the optimization of amino acid classes in a journal paper that is currently in preparation.

Conclusion: The described method is to our best knowledge the only existing result on MI confidence interval estimation for high dimensional random variables when only an i.i.d. sample is given and no further knowledge on the joint distribution, and we think that it has many applications in wide variety of fields, especially in molecular biology.

Acknowledgements: The authors would like to thank the DFG for supporting their research with SPP1395 in the projects HU634_7 and STI155_3.

References:

[1] T. M. Cover and J. A. Thomas, Elements of Information Theory, 2nd ed. New York: Wiley, 2006.

[2] S. P. Lloyd, “Least squares quantization in PCM”, IEEE Trans. Inform. Theory, vol. 28, no. 2, pp. 129–137, Mar. 1982.

[3] A. G. Stefani, J. B. Huber, C. Jardin and H.Sticht, “Confidence intervals for the mutual information,” submitted to ISIT 2013, available at http://arxiv.org/abs/1301.5942.

ARE CANCER CELLS GOOD PLAYERS?

Andrzej Świerniak and Michał Krześlak

Institute of Automatic Control, Silesian University of Technology,
Akademicka 16, 44-100 Gliwice, Poland,
andrzej.swierniak@polsl.pl

ABSTRACT

The theory of games provides a very powerful tool for analysis of processes in which decision-making plays an important role. From its very beginning it was mainly applied in economics and econometrics, and soon it was also used successfully to solve problems in behavioural and social sciences, control and process engineering, and military situations. Apart from these applications, new perspectives in biology were introduced by John Maynard Smith and George Price [1, 2]. Their ideas linked the mathematical tools of game theory with Darwinian adaptation and species evolution, and initiated a new branch in decision-making mathematics called evolutionary game theory (EGT). This new approach differs from standard game theory by incorporating rational decision-making by the competing players, strategies are treated as phenotypes of individuals in the population acquired through evolution, and payoffs measure a change in the degree of fitness resulting from interactions of the individuals representing different phenotypes. EGT is based on the assumption of perfect mixing inside the population and interaction of each pair of strategies at one time. To overcome this simplification, evolutionary games have been transferred into spatial lattices by application of cellular automata techniques where an additional important factor, namely spatial allocation, is included. Although the origin can be found in the pioneering works of von Neuman [3], Nowak and May have usually been granted the name of the fathers of spatial evolutionary games theory (SEGT) [4]. As mentioned earlier, EGT has been used in biology to predict the survival of different phenotypes in a population. To check how and when a population becomes stable it is necessary to simulate phenotype interactions among generations according to a payoff matrix. One way in which the dynamics of transients from an initial to new stable states could be studied is the use of replicator dynamics equations (RD) [5].To our knowledge, the first work in which evolutionary game theory was used to model the interaction behaviour of tumour cells was presented by Tomlinson and Bodmer [6], who proposed the model where one of the phenotypes attempts to gain an advantage by producing cytotoxic substances. The results show that actively harming neighbouring cells may lead to dominance of the local population by the tumour cells. This study triggered a series of other papers, and below we overview the features of the models discussed in these publications and present the main results. We append to this analysis our results obtained by SEGT and RD tools if absent in the original study.

We review a quite large volume of literature concerning mathematical modelling of processes related to carcinogenesis and the growth of cancer cell populations based on the theory of evolutionary games. This review, although partly idiosyncratic, covers such major areas of cancer-related phenomena as production of cytotoxins, avoidance of apoptosis, production of growth factors, motility and invasion, and intra- and extracellular signaling. We discuss the results of other authors and *append* to them some additional results of our own simulations dealing with the possible dynamics and/or spatial distribution of the processes discussed. Moreover we present also some our original results.

Keywords: evolutionary games, cancer, replication dynamics, cellular signalling

Literature

[1] Maynard Smith J. Evolution and the theory of games. Cambridge :Cambridge University Press (1982).

[2] Maynard Smith J., Price G.R. The Logic of Animal Conflict. Nature, 246 (1973), 16-18.

[3] von Neuman J. Theory of self reproducing automata. University of Illinois Press, (1966).

[4] Nowak M.A., May R.M. Evolutionary games and spatial chaos. Nature, 18 (1992), 826-829

[5] Hofbauer J., Sigmund K. Evolutionary game dynamics. Bull. Amer. Math. Soc. 40 (2003), 479-519

[6] Tomlinson I.P.M., Bodmer W.F. Modeling the consequences of interactions between tumour cells. British Journal of Cancer, 75 (1997), 157-180.

IMAGE PROCESSING BASED CLASSIFIER FOR AUTOMATED PREDICTION OF OVARIAN CANCER RECURRENCE

Francesco Tabaro[1], Pekka Ruusuvuori[1,2], Yuexin Liu[3], Wei Zhang[3], Matti Nykter[1]

[1]Computational Biology, Institute of Biomedical Technology, University of Tampere, Finland,
[2]Department of Signal Processing, Tampere University of Technology, Finland
[3]Departments of Pathology, The University of Texas MD Anderson Cancer Center, Houston, Texas, United States of America
francesco.tabaro@uta.fi

ABSTRACT

Ovarian cancer (OvCa) is the eighth most deadly tumor in the United States and despite a general low incidence worldwide in the Scandinavian countries it has an higher incidence. Primary tumors are, generally, treated in a surgical way followed by a variable number of chemotherapic cycles based on the diagnosed stage (FIGO system). Even with the pharmacological treatments recurrent tumors appear in more than 50% of patients with high grade OvCa. These are treated in a pharmacological way with different chemotherapic drugs such as cisplatin, paclitaxel and doxorubicin. Unfortunately, not all of them are sensitive to such treatments. Thus, it is a major clinical challenge to predict whether a recurrent form of OvCa is resistant to platinum containing drugs or not. Here we use digital image processing and Random Forest classifier to predict recurrent forms of OvCa in a automated fashion from histological tissue data.

The dataset is made up of 2444 high resolution histological images from 488 different patients obtained from The Cancer Genome Atlas Project enclosed with clinical data about each patient. Each image is processed to extract features in an automated fashion. The basic workflow follows implies filtering, k-means clustering, watershed segmentation and outliers removal. After these steps statistical analysis of the detected cell nuclei is performed to build the feature matrix input of the classifier.

The computed feature matrix together with platinum sensitivity status coming from the clinical data are used to train a Random Forest classifier. In order to evaluate the predictive power of our features set we split the dataset in two parts: training set (60% of the whole data) and testing set (40% of data). In the initial testing of our pipeline we got 73% of out-of-bag classification accuracy during the training phase, and 63% on a blind test on the testing set. 100-fold cross validation lead to similar results: 69% of out-of-bag accuracy in the training phase and 70% in the blind test.

These results show that our feature set from histological data has predictive power of OvCa recurrence even if more refinement work has still to be done. Ongoing work aims to improve the image processing part and extend the classification to incorporate features from molecular data, including expression data and genomic alterations. These improvements and extensions should lead to even higher prediction accuracy.

MRMQUANT – A FLEXIBLE MRM-DATA ANALYSIS TOOL FOR METABOLOMICS AND FLUXOMICS

Max von Haugwitz, Nicole Paczia, Wolfgang Wiechert and Katharina Nöh

IBG-1: Biotechnology & JARA-HPC, Forschungszentrum Jülich GmbH,
52425 Jülich, Germany
m.von.haugwitz@fz-juelich.de

ABSTRACT

Isotope labeling experiments (ILE) are a well-established tool for metabolomics and fluxomics studies aiming at the absolute quantification of central metabolic concentrations and reaction rates in living cells [1]. Low intracellular metabolite concentrations and complex biological matrices require highly selective and sensitive experimental methods, e.g., liquid chromatography combined with Multiple Reaction Mode mass spectrometry (MRM-based LC-MS). Data analysis of complex chromatograms is a major bottleneck of the evaluation pipeline requiring the development of effective and accurate evaluation tools.

Most software packages for analysis of LC-MS data are either not compatible to MRM data or focus solely on proteomics applications [2]. Typically, integration results are unsatisfactory and imply time-consuming manual adjustments. To circumvent such limitations, we propose a white-box approach using advanced processing techniques from signal processing and pattern recognition combined in a single evaluation workflow. In particular, the similarity of mass chromatograms typically found in ILE-generated data, is exploited. This enables speeding up the evaluation process while keeping confidence in the analysis results. A graphical user interface allows for user intervention at important steps and visualizes various quality measures.

We benchmarked our analysis workflow with a ^{13}C-ILE data set consisting of approx. 2.000 chromatograms. In 95% of the cases a deviation of less than 2% compared to an expert operator in terms of ^{13}C labeling fractions was found. In-depth analysis of the deviates revealed two sources for discrepancies: (1) software integration errors (4%) and (2) non-decidable cases in which rendering either the human or software solution as correct or wrong is hardly possible. The latter case was investigated using a data set with chromatograms of different complexity. A user study with 10 participants showed that (1) for complicated chromatograms a scatter of expert solutions emerges and (2) solutions of MRMQuant are well comparable to those of human operators.

All in all, results generated by MRMQuant are promising, providing a first step towards a more reliable data processing toolkit. The workload of human operators is strongly reduced while confidence into the analysis results is kept. However, the study demonstrates that evaluation errors are ever-possible (for human and software), and post-integration error checking remains time-consuming. Future work focuses on improving the quality indicators and proposing reliable quality measures. Machine learning techniques are evaluated to enhance the high-throughput processing capabilities of MRMQuant.

[1] Rühl et al., Biotechnol Bioeng, 2011, 109(3):763-71.
[2] Wong et al., Anal Chem., 2012, 84(1): 470-4.

A DETERMINISTIC METHOD FOR INFERENCE IN STOCHASTIC MODELS

Christoph Zimmer[1,2] and Sven Sahle[1]

[1] BioQuant, University of Heidelberg, Germany,
[2] HGS MathComp, University of Heidelberg, Germany
christoph.zimmer@bioquant.uni-heidelberg.de

ABSTRACT

Parameter estimation is very important for the analysis of models in Systems Biology. Stochastic models are of increasing importance. We show a fast and efficient method for parameter estimation in stochastic models. The method approximates the stochastic model on relatively small time intervals by a system of ordinary differential equations (ODE). Every measurement is used for readjusting the approximation. It is shown that the method works well even in partially observed systems which behave qualitatively different in stochastic modeling than in ODE modeling. As the method is based on ODEs it allows from computational point of view to tackle systems as large as those tackled in deterministic modeling.

Introduction: Computational modeling is a central approach in Systems Biology for studying increasingly complex biochemical systems. Progress in experimental techniques, e.g. the possibility to measure small numbers of molecules in single cells [1] highlights the need for stochastic modeling approaches. Simulation methods for stochastic processes are being developed for decades since [2] and nowadays exist with a lot of variants [3]. The development of parameter estimation methods for stochastic models however has recently started. We present a recently developed approach for parameter estimation in stochastic models [4] and its performance on a stochastic model of a genetic toggle switch [5].Related work uses finite state projection [6] to tackle the chemical master equation or moment closure methods [7]. Stochastic simulations in combination with density estimation methods can be found in [8]. The common challenge is the size of the state space.

Method: The MSS method uses one single realization of an intrinsic stochastic time course as input data. The data is recorded at discrete times. The method approximates the stochastic model on relatively small time steps with a system of ordinary differential equations (ODE). Every new measurement is used for the readjustment of the approximation. This allows for capturing the stochastic dynamics. Unobservable species are treated as optimization variables. As the method uses a deterministic objective function it can be optimized using global, Bayesian or derivative based optimization methods.

Results: The first test model is an Immigration-Death model in steady state. Using conventional ODE methods for parameter estimation the model is structurally non-identifiable. The MSS method is able to exploit the information in the stochastic fluctuations and resolve the structural nonidentifiability. The second model is a stochastic model of a genetic toggle switch. This model of the genetic toggle switch has a stable steady state in ODE modeling but shows a switching behavior in stochastic modeling. As data input for the method a single recording of one of the species at discrete time points is used. A reliable test of the MSS method's performance must be independent of the intrinsic stochasticity of a single realization. To address this point, the estimation procedure is repeated 50 times which yields 50 estimates. Statistics over these will be shown. The repeating of the estimation procedure is only done for testing purposes. For the estimation of parameters from experimental data one single recording is enough. Results on a genetic toggle switch show that the method is able to estimate the parameters very fast and with acceptable accuracy. Some larger confidence intervals are due to identifiability problems.

Conclusion: We show a method for parameter estimation in stochastic models based on an approximation with ODEs on a relatively short time scale. This performance of the method is demonstrated by estimating successfully the parameters in a stochastic model of a genetic toggle switch which has in stochastic modeling a qualitatively different dynamical behavior than in ODE modeling. The advantage of the method is high speed as no stochastic simulations or solutions of the chemical master equation are needed.

References

[1] A. Raj and A. van Oudenaarden. *Annu. Rev. Biophys.*, 38:255–270, 2009.
[2] D.T. Gillespie. *Journal of Computational Physics*, 22 (4):403–434, 1976.
[3] J. Pahle. *Briefings in Bioinformatics*, 10 (1):53–64, 2009.
[4] C. Zimmer and S. Sahle. *Journal of Computer Science & Systems Biology*, 6:011–021, 2012.
[5] T. S. Gardner, C. R. Cantor, and J. J. Collins. *Letters to Nature*, 403:339–342, 2000.
[6] B. Munsky and M. Khammash, *IET Systems Biology*, 4 (6):356–366, 2010.
[7] C.S. Gillespie and A. Golightly. *Applied Statistics*, 59 (2):341–357, 2009.
[8] T. Tian, S. Xu, J. Gao, and K. Burrage. *Bioinformatics*,23 (1):84–91, 2007.

DERIVATIVE PROCESSES FOR MODELLING METABOLIC FLUXES

Justina Zurauskiene[1], Paul Kirk[1], Tom Thorne[1], John Pinney[1] and Michael Stumpf[1]

[1]Theoretical Systems Biology group, Imperial College London, South Kensington Campus, London SW7 2AZ, UK
j.norkunaite@imperial.ac.uk

ABSTRACT

One of the challenging questions in modelling biological systems is to characterize the functional forms of the processes that control and orchestrate molecular and cellular processes. Recently proposed methods for the analysis of metabolic pathways, for example dynamic flux estimation, can only provide estimates of the underlying fluxesin a point-wise fashion at discrete time-points (mostly, in fact, just a single time-point) but fail to capture the complete temporal behaviour. In order to describe the dynamic variation of the fluxes we additionally require the assumption of specific functional forms that can capture the temporal behaviour. Here we propose a novel approach to modelling metabolic fluxes: derivative processes that are based on Multiple-output Gaussian processes (MGPs), which are a flexible nonparametric Bayesian modelling technique. Our derivative process approach does not require detailed knowledge of the dynamics of regulatory/metabolic pathways or corresponding ODE models.

Our approach allows us to characterize the temporal behaviour of metabolic fluxes from time course data. Because the derivative of a Gaussian process is itself a Gaussian process we can readily link metabolite concentrations to metabolic fluxes and vice versa. Here we discuss how this can be implemented in an MGP framework and illustrate its application to simple metabolic models, including nitrogen metabolism in Escherichia coli.

Tampereen teknillinen yliopisto
PL 527
33101 Tampere

Tampere University of Technology
P.O.B. 527
FI-33101 Tampere, Finland

ISBN 978-952-15-3091-3 (printed)
ISBN 978-952-15-3092-0 (PDF)
ISSN 1456-2774