MLSB14

The eighth International Workshop on Machine Learning in Systems Biology

6-7 September 2014

Strasbourg, France

Editors:
Florence d'Alché-Buc
Pierre Geurts

Cover design: Vincent Botta

Contents

MLSB14, the eighth International Workshop on Machine Learning in Systems Biology takes place in Strasbourg on September 6-7, 2014.
The aim of this workshop is to contribute to the cross-fertilization between the research in machine learning methods and their applications to systems biology (i.e., complex biological and medical questions) by bringing together method developers and experimentalists.

Conference Chairs

Florence d'Alché-Buc, IBISC, Université d'Evry, Genopole & LRI, Université Paris Sud, France
Pierre Geurts, University of Liège, Institut Montefiore, Liège, Belgium

Organizing Support

Markus Heinonen (University of Evry, France)
Vân Anh Huynh-Thu (University of Edinburgh, UK)
Nizar Touleimat (Centre National de Genotypage, Genopole, Evry)

Programme Committee

Florence d'Alché-Buc (University of Evry, France)
Chloé-Agathe Azencott (CBIO, Mines ParisTech, Institut Curie, INSERM, France)
Karsten Borgwardt (ETH Zürich, Switzerland)
Sašo Džeroski (Jožef Stefan Institute, Slovenia)
Mohamed Elati (Universite d'Evry, France)
Pierre Geurts (University of Liège, Belgium)
Markus Heinonen (University of Evry, France)
Vân Anh Huynh-Thu (University of Edinburgh, UK)
Ross King (Aberystwyth University, UK)
Stefan Kramer (University of Mainz, Germany)
Yves Moreau (Katholieke Universiteit Leuven, Belgium)
Sach Mukherjee (Netherlands Cancer Institute, Netherlands)
Mahesan Niranjan (University of Southampton, UK)
Uwe Ohler (Duke University, USA)
John Pinney (Imperial College London, UK)
Simon Rogers (University of Glasgow, UK)
Juho Rousu (Aalto University, Finland)
Yvan Saeys (University of Ghent, Belgium)
Peter Sykacek (BOKU University, Austria)
Nizar Touleimat (Centre National de Genotypage, CEA, France)
Koji Tsuda (National Institute of Advanced Industrial Science and Technology, Japan)
Jean-Philippe Vert (Ecole des Mines de Paris, France)
Filip Zelezny (Czech Technical University in Prague, Czech Republic)

Sponsors

The organizers gratefully acknowledge the following sponsors who have provided financial support or/and precious help for the organization of the workshop.

IBISC laboratory, Evry, France

University of Evry, France

Genopole, France

GIGA research, University of Liège, Belgium

Schedule

Saturday, 6th September

9am-10:30am Session 1

Introduction

Invited talk: Pierre Baldi, UCI University. *Carbon-Based Computing Vs Silicon-Based Computing: A New Theory of Circadian Rhythms* (p11)

Daniel Trejo-Banos. *Structural inference in oscillatory networks: a case study of the Arabidopsis Thaliana circadian clock* (p65)

10:30am-10:45am Coffee break

10:45am-12pm Session 2

Xin Liu. *Parameter Estimation in Computational Biology by Approximate Bayesian Computation coupled with Sensitivity Analysis* (p48)

Vân Anh Huynh-Thu. *A hybrid approach for the inference and modelling of gene regulatory networks* (p39)

Karel Jalovec. *Classification of metagenomic samples using discriminative DNA superstrings* (short talk) (p44)

12pm-1:30pm Lunch

1:30pm-3:30pm Session 3

Invited talk: Nicola Segata, University of Trento. *Machine learning challenges in computational meta'omics* (p16)

Eugen Bauer. *Metabolic Meta-Reconstruction and Community Modeling of Intestinal Microbes* (p23)

Aalt-Jan van Dijk. *Interspecies Association Mapping: connecting phenotypes to sequence regions across species* (p69)

3:30pm-4:30pm Poster session

4:30pm-6pm Session 4

Invited talk: Anne-Laure Boulesteix, LMU Munich. *Statistical testing and variability in real-data-based benchmark experiments for supervised learning methods* (p13)

Adrien Dessy. *Computationally Efficient Test for Gene Set Dysregulation* (short talk) (p31)

Sohan Seth. *Differential analysis of whole-genome shotgun sequences* (short talk) (p61)

Sunday, 7th September

9am-10:30am Session 5

Invited talk: Jean-Loup Faulon, CNRS. *Using Machine Learning in Synthetic Biology: The Design-Build-Test and Learn cycle* (p14)

Tom Mayo. *M3D: a kernel-based test for shape changes in methylation profiles* (p52)

10:30am-10:45am Coffee break

10:45am-12pm Session 6

Pooya Zakeri. *Application of Geometric Kernel Data Fusion in Protein Fold Recognition and Protein Sub-nuclear Localization* (p73)

Yawwani Gunawardana. *Outlier-Detecting Support Vector Regression for Modelling at the Transcriptome-Proteome Interface* (p35)

Olivier Poirion. *Structuration of the bacterial replicon space* (short talk) (p57)

12pm-1:30pm Lunch

1:30pm-3:30pm Session 7

Invited talk: Karsten Borgwardt, ETH Zürich. *Machine Learning for Personalized Medicine* (p12)

Anna Cichonska. *Meta-analysis of Genome-Wide Association Studies with Multivariate Traits* (p27)

Roland Barriot. *Semi-automatic Validation of Genome-wide Reassembled Systems by Gene Prioritization through Weighted Data Fusion* (p19)

3:30pm-4:20pm Poster session

4:20pm-5:20pm Session 8

Invited talk: Katheen Marchal, KU Leuven & U Ghent. *Network-based data-integration: applications to clonal systems* (p15)

Closing remarks

Invited talks

List of invited talks

1. *Carbon-Based Computing Vs Silicon-Based Computing: A New Theory of Circadian Rhythms.*
   Pierre Baldi. *p11*

2. *Machine Learning for Personalized Medicine.*
   Karsten Borgwardt. *p12*

3. *Statistical testing and variability in real-data-based benchmark experiments for supervised learning methods.*
   Anne-Laure Boulesteix. *p13*

4. *Using Machine Learning in Synthetic Biology: The Design-Build-Test and Learn cycle.*
   Jean-Loup Faulon. *p14*

5. *Network-based data-integration: applications to clonal systems.*
   Kathleen Marchal. *p15*

6. *Machine learning challenges in computational meta'omics.*
   Nicola Segata. *p16*

Carbon-Based Computing Vs Silicon-Based Computing: A New Theory of Circadian Rhythms

Pierre Baldi
UCI University, California, USA

Abstract
Carbon-based and silicon-based computing systems are very different. One key difference is the pervasive presence of circadian rhythms in living systems at multiple levels. At the molecular level, circadian rhythms are regulated by a central clock consisting of a key negative transcription-translation feedback loop involving a dozen of genes. However, integrative systems biology analyses of high-throughput transcriptomic and metabolomic data reveal that roughly 10% of genes or metabolites oscillate in a circadian manner in any given cell or tissue. Furthermore, when data is aggregated across different tissues and genetic or environmental conditions, the overlap in circadian species beyond the core clock is very small. Thus a large fraction of molecular species in the cell is capable of oscillating in a circadian manner under some set of conditions. We will present a novel theory of circadian rhythms to explain these puzzling findings. In this theory, molecular networks are viewed as networks of coupled-oscillators sculpted by 3.5 billion years of evolution. Under a given set of genetic and environmental conditions, a cell can reprogram itself and select its own subset of oscillatory species out of a vast repertoire. The oscillating species provide a physiological signature of the state of the cell.

- V. R. Patel, K. Eckel-Mahan, P. Sassone-Corsi, and P. Baldi. *How Pervasive Are Circadian Oscillations?* Trends in Cell Biology, in press, DOI:10.1016/j.tcb.2014.04.005, (2014).
- K. L. Eckel-Mahan1, V. R. Patel, S. de Mateo, N. J. Ceglia, S. Sahar, S. Dilag, Kenneth A. Dyar, R. Orozco-Solis, P. Baldi, and P. Sassone-Corsi. *Reprogramming of the Circadian Clock by Nutritional Challenge.* Cell, 155, 7, 1464-1478, (2013).
- V. Patel, K. Eckel Mahan, P. Sassone-Corsi, and P. Baldi. *CircadiOmics: Integrating Circadian Genomics, Transcriptomics, Proteomics, and Metabolomics.* Nature Methods, 9, 8, 772-773, (2012).
- K. L. Eckel-Mahan, V. R. Patel, K. S.Vignola, R. P. Mohney, P. Baldi, and P. Sassone-Corsi. *Coordination of Metabolome and Transcriptome by the Circadian Clock.* PNAS, 109 (14) 5541-5546, (2012)

Machine Learning for Personalized Medicine

Karsten Borgwardt
Department of Biosystems Science and Engineering (D-BSSE), ETH Zürich, Switzerland

Abstract
Over the last decade, enormous progress has been made on recording the health state of an individual patient down to the molecular level of gene activity and genomic information - even sequencing a patient's genome for less than 1000 dollars is within reach. However, the ultimate hope to use all this information for personalized medicine, that is to tailor medical treatment to the needs of an individual, remains largely unfulfilled. To turn the vision of personalized medicine into reality, many methodological problems remain to be solved: there is a lack of methods that allow us to gain a causal understanding of the underlying disease mechanisms, including gene-gene and gene-environment interactions. Similarly, there is an urgent need for integration of the heterogeneous patient data currently available, for improved and robust biomarker discovery for disease diagnosis, prognosis and therapy outcome prediction. The field of machine learning, which tries to detect patterns, rules and statistical dependencies in large datasets, has also witnessed dramatic progress over the last decade and has had a profound impact on the Internet. Amongst others, advanced methods for high-dimensional feature selection, causality inference, and data integration have been developed or are topics of current research. These techniques address many of the key methodological challenges that personalized medicine faces today and keep it from rising to the next level. In this talk, we will describe the challenges and opportunities for machine learning in personalized medicine and we will present our recent research results in this direction.

Statistical testing and variability in real-data-based benchmark experiments for supervised learning methods

Anne-Laure Boulesteix
Department of Computational Molecular Medecine, IBE, Ludwig-Maximilians Universität München, Germany

Abstract
Resampling-based methods such as, e.g., k-fold cross-validation or repeated splitting into training and test sets are routinely used in the context of supervised statistical learning to assess the prediction performance of prediction methods. In this talk, I discuss two important issues related to the use of such methods: the design of resampling-based benchmark experiments from the perspective of statistical testing and the variability of resampling-based procedures for the choice of tuning parameters.

The first part of the talk deals with benchmark experiments that aim at comparing the performance of different algorithms and presents a statistical framework for hypothesis testing in real data comparison studies. In computational literature, most abstracts of articles presenting new supervised learning methods end with a sentence like "our method performed better than existing methods on real data sets", e.g. in terms of error rate. However, these claims are often not based on proper statistical inference and, if statistical hypothesis tests are performed, the tested hypothesis is not clearly defined and poor attention is paid to the type I and type II error. We propose a proper statistical framework for hypothesis tests comparing the performance of supervised learning methods based on several real data sets with unknown underlying distributions. After giving a statistical interpretation of ad-hoc paired t-tests commonly performed in practice, we devote special attention to power issues and outline a simple method to determine the number of data sets to be included in a comparison study to reach an adequate power. As an extension of this testing framework, we will also consider regression models considering the relative performance of prediction methods as a dependent variable and data sets' characteristics as independent variables, with the long-term aim to guide the choice of the prediction method depending on the type of data set at hand. The methods are illustrated through an application of PLS-based classification to 50 microarray gene expression data sets.

The second part of the talk is devoted to the problem of the choice of tuning parameters - for instance the number of PLS components - based on resampling methods. I will again show results of a PLS-based classification method applied to microarray data sets. These results empirically demonstrate the instability induced by the random character of resampling-based procedures. As suggested by this study and substantiated by theoretical results, we recommend to perform a large number of resampling iterations whenever computationally possible to yield a better stability.

Using Machine Learning in Synthetic Biology: The Design-Build-Test and Learn cycle

Jean-Loup Faulon
IBSS, CNRS, Genopole, University of Evry, France

Abstract
Synthetic biology and metabolic engineering have succeeded in the biosynthesis of numerous commodity or high value compounds. Yet, the choice of pathways and enzymes used for such successful applications was many times made ad hoc, or required expert knowledge of the specific biochemical reactions. In order to rationalize this process we have developed the computer-aided design (CAD) tool RetroPath [1] that explores and enumerates metabolic pathways connecting the endogenous metabolites of a chassis cell to a target compound. Namely, our tool queries for target activities the list of enzymes found in metabolic databases based on their annotated and predicted activities based on a tensor product kernel [2]. Next, it ranks pathways based on the predicted efficiency of the available enzymes, the toxicity of the intermediate metabolites and the calculated maximum product flux. As an illustration of the power of rational design, RetroPath compiled the top-ranking pathways producing the flavonoid pinocembrin (a antibacterials targeting Staphylococcus aureus), narrowing down a list of nine million possible enzyme combinations to a number that could be easily assembled and tested. We next constructed the top-ranked enzyme combinations, four of which displayed significant yields. One round of metabolic network optimization based on RetroPath output further increased pinocembrin titers 17-fold [3]. In total, 12 out of the 13 enzymes tested in this work displayed a performance that was in accordance with its predicted score. These results validate the ranking function of our CAD tool, and open the way to its utilization in the biosynthesis of novel compounds.

1. Carbonell P, Parutto P, Herisson J, Pandit S.B, Faulon J.L. *XTMS: pathway design in an eXTended metabolic space.* Nucleic Acids Res. in press 2014, [PMID: 24792156].
2. Faulon J.L., Misra M., Martin S., Sale, K., Sapra R. *Genome Scale Enzyme-metabolites and Drug-Target interaction predictions using the signature molecular descriptor.* Bioinformatics, 24, 225-233, 2008.
3. Fernandez-Castane A, Feher T, Carbonell P, Pauthenier C, Faulon J.L. *Computer-aided design for metabolic engineering.* J Biotechnol. in press 2014, [PMID: 24704607].

Network-based data-integration: applications to clonal systems

Kathleen Marchal
Depts. Plant Biotechnology and Bioinformatics; Information Technology (INTEC, iMINDS), U.Ghent;
Dept. of Microbial and Molecular Systems, K.U.Leuven, Belgium

Abstract
In the last 10 years huge efforts have been made to infer interaction networks from large scale omics data. Networks have been inferred at different molecular levels (transcriptional, metabolic, protein interaction, signaling), allowing us to compile for a model organism 'integrated networks' in which nodes represent molecular entities (genes, proteins etc) and edges the interactions between those entities. Such networks that span different molecular levels, despite being noisy and static and thus overconnected, do comprehensively summarize all available (reliable and less reliable) molecular knowledge on an organism of interest. In this presentation we will show how the information contained within those networks can be exploited to guide experimentalists with the interpretation of their own in house generated omics experiments or how some data-integration efforts depend on network-based guidance (such as for the interpretation of genomic variations in clonal systems).

Machine learning challenges in computational meta'omics

Nicola Segata
Centre for Integrative Biology, University of Trento, Italy

Abstract
The study of the microbial diversity with sequencing-based cultivation-free approaches (metagenomics) is currently revolutionizing our understanding of the biology associated with many natural systems including the human body. Outnumbering our own cells 10 to 1, the cloud of microbial organisms (microbiome) living in symbiosis with our body has, in fact, a profound effect on human health and is co-responsible of several complex diseases when the microbiome/host equilibrium is broken. Following the recent biotechnological revolution that enabled shotgun metagenomics, the research community focused primarily on developing computational tools to extract meaningful taxonomic [1], phylogenetic [2], and functional profiles [3] of the microbiome from the large amount of raw-data produced. These methods are now providing reliable quantitative snapshots of the microbiome that can thus be used for advanced learning tasks of clinical relevance including biomarker discovery [4] and subtype identification. However, specific characteristics of metagenomic data are affecting the effectiveness of such tasks and the availability of complementary meta'omic techniques (e.g. metatranscriptomics) is posing novel issues [5]. In this talk, I will discuss the machine learning challenges that the field is facing and will present some of the most relevant learning tasks in the study of the human-associated microbiome.

1. Segata, N., Waldron, L., Ballarini, A., Narasimhan, V., Jousson, O., & Huttenhower, C. (2012). Metagenomic microbial community profiling using unique clade-specific marker genes. Nature Methods, 9, 811-814.
2. Segata, N., Börnigen, D., Morgan, X., & Huttenhower, C. (2013). PhyloPhlAn is a new method for improved phylogenetic and taxonomic placement of microbes. Nature Communications, 4, 2304.
3. Abubucker, S., Segata, N., Goll, J., Schubert, A., Izard, J., Cantarel, B. L., ... Huttenhower, C. (2012). Metabolic Reconstruction for Metagenomic Data and Its Application to the Human Microbiome. PLoS Computational Biology, 8(6), e1002358.
4. Segata, N., Izard, J., Waldron, L., Gevers, D., Miropolsky, L., Garrett, W. S., & Huttenhower, C. (2011). Metagenomic Biomarker Discovery and Explanation. Genome Biology, 12, R60.
5. Segata, N., Boernigen, D., Tickle, T. L., Morgan, X., Garrett, W. S., & Huttenhower, C. (2013). Computational meta'omics for microbial community studies. Molecular Systems Biology, 9(666), 1-15.

Extended abstracts

List of abstracts

Semi-automatic Validation of Genome-wide Reassembled Systems by Gene Prioritization through Weighted Data Fusion

Roland Barriot, Petra Langendijk-Genevaux, Yves Quentin and Gwennaele Fichant

Université de Toulouse; UPS; Laboratoire de Microbiologie et Génétique Moléculaires; F-31000 Toulouse ; France.
Centre National de la Recherche Scientifique; LMGM; F-31000 Toulouse; France

Abstract: ATP-Binding Cassette (ABC) systems constitute a major super-family of systems present in all kingdoms of life, composed, in prokaryotic genomes, of two to five partners. We previously developed an annotation and reconstruction pipeline and maintain an online database ABCdb. The pace of new complete genomes releases requires tools to assist experts in the validation of the reassembled systems. Here, we present a method for quality assessment of the reconstruction inspired by gene prioritization through genomic data fusion. The main innovations are (i) the weighing of the data sources used for prioritization and (ii) a genome wide approach for assemblies validation.

Background

ATP-Binding Cassette (ABC) systems constitute a major super-family of proteins present in all kingdoms of life. They are mainly involved in the active transport of a large variety of compounds such as amino-acids, sugars, metal ions, and so on. In prokaryotic genomes, they are composed of two to five partners and one major difficulty in the analysis of these systems arises from the very high level of paralogy encountered. For example, *Bradyrhizobium japonicum* strain USDA110 harbors 696 genes involved in 227 ABC systems. Previous work on these systems lead to an automated pipeline for their assembly into functional systems mainly based on genome sequence. Afterwards, the systems are classified into subfamilies through sequence similarity to manually expertized systems in whole genomes. Both expertized and reconstructed systems are available in a dedicated online database ABCdb[1]. The pace of new complete prokaryotic genome releases (>1.5/day, source: GOLD[2]) requires tools to assist experts in the validation of the reassembled systems.

Here, we present a strategy for the quality assessment of the reconstruction inspired by gene prioritization through genomic data fusion[3]. The main innovations are (i) the weighing of the data sources used for prioritization and (ii) a genome wide approach for assemblies validation. The strategy is to take advantage of data sources external to genome sequence to challenge and evaluate the quality of the reassembled systems. For instance, genes coding for proteins involved in the same system are expected to be co-regulated, and thus should exhibit co-expression. They also should be either all present or absent in a genome, and thus should exhibit similar phylogenetic profiles. In gene prioritization, candidate genes are scored according to their dissimilarity to a set of training genes using dissimilarity matrices derived from various data sources. As a result, a partner of a well reconstructed system should rank first among all the ABC coding genes of a genome when the training set corresponds to its other partners.

By simultaneously considering all the systems of one genome by a graph summarizing the best candidate for each system, we are able to validate automatically most of the reconstructed systems and pinpoint a few remaining more complex scenarios. The strategy is illustrated on the well annotated and expertized ABC systems of *Escherichia coli* K-12 MG1655.

Data Sources

In the following paragraphs, we describe the data sources used and how pairwise gene dissimilarity matrices were derived from them.

Expression data. Raw hybridization data were downloaded from GEO[4] and normalized in R/Bioconductor[5]. For a first normalization, the Robust Multiarray Averaging (RMA) algorithm was used (R/Affy package[6]). Then a second was done as advised by Oti and colleagues[7], *i.e.* each gene expression value was divided by the median expression value of all considered genes in the sample. To derive a gene pairwise dissimilarity matrix, profiles dissimilarity was computed as one minus the Pearson correlation coefficient divided by 2 to obtain a value between 0 and 1: $(1 - r) / 2$.

Annotations. For annotations, we used the Gene Ontology[8] and the gene product associations obtained from GOA[9]. After benchmarking various pairwise GO term similarity indices and various combinations for deriving gene pairwise similarity indices (see [10] for more details), we selected to measure GO term similarity with the

information content of the Maximum Information content Common Ancestor (MICA). The information content (IC) of a GO term is defined as -log(freq(term)) with freq(term) being the fraction of genes annotated with that term in the genome. Then for a pair of terms, the one that exhibits the maximal IC over the common ancestors of the two terms in the GO hierarchy is retained. Then, for a pair of genes, both possibly annotated with multiple GO terms, the maximum of all pairwise term similarities is retained. The final dissimilarity is obtain by 1 – similarity.

Interactions. For functional interactions, we used the STRING database[11]. The whole STRING release 9.05 was downloaded. The combined protein association scores were filtered to keep only associations with a score of at least 600. The gene pairwise dissimilarity is computed as the shortest path length between the two genes in the filtered STRING graph. The dissimilarity is afterwards normalized between 0 and 1.

Phylogenetic profiles. For phylogenetic profiles, we used a method commonly referred to as genomic context methods[12] or Gene Neighbors (GN). A set of reference genomes is selected based on the core genome and evolutionary distance to the genome investigated as well as between reference genomes. Orthologs 1:1 were inferred based on the rule proposed in [13]. It specifies that two sequences *a* and *b* from genomes *A* and *B* are orthologs 1:1 if they are bidirectional best hits (BBH) and there is no other sequence in either genome *A* or *B* having a better alignment score with *a* or *b*. Then, for each pair of genes, the dissimilarity is computed as the joint probability that the distances between their orthologs 1:1 in the reference genomes are smaller or equal to their observed distances. $X = \Pi\, P(Di \leq di) = \Pi\, 2d_i / (N_i\text{-}1)$, with d_i the observed distance (in genes) in the i^{th} genome and N_i the size of i^{th} genome. As the number of genomes for which orthologs are found depends on the considered gene pair, a normalization step is required, and thus performed as advised in [12]. Roughly, it consists in replacing the probabilities by z-scores (distribution of probabilities across one row) and then averaging the obtained matrix with its transpose to have a symmetric dissimilarity matrix.

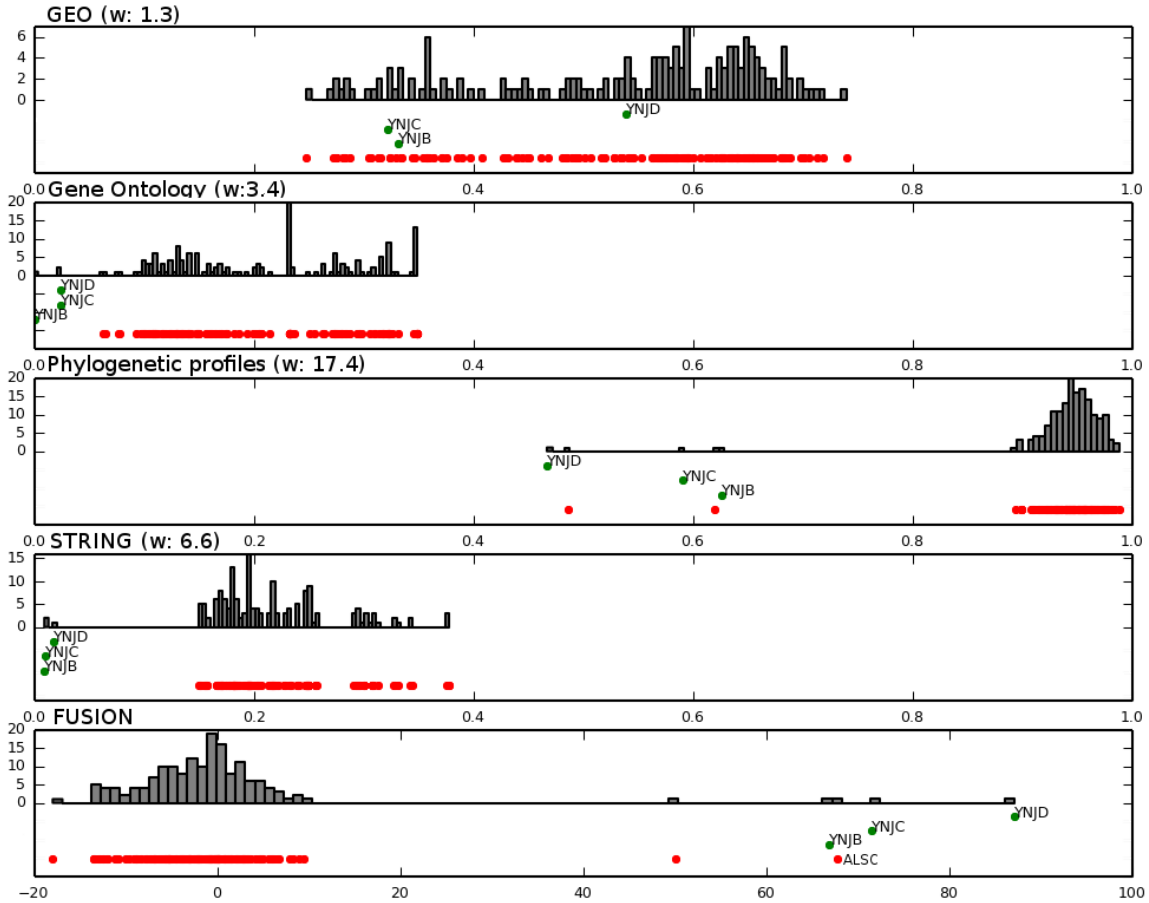


Figure 1. Prioritizations of ABC coding genes of *E. coli* with respect to the YNJD assembly genes with 4 data sources: expression data (GEO), annotations (Gene Ontology), phylogenetic profiles and interaction (STRING). The top ranking genes are at the left (lowest dissimilarity).The last plot correspond to the weighted fusion obtained by linear discriminant analysis. The top ranking genes are on the right. The weight attributed to each data source is indicated between parentheses. Histograms: distribution of the prioritized genes. Green dots indicate genes from the reassembled ABC system. Red dots indicate genes from other ABC systems.

Gene Prioritization through Weighted Data Fusion

The process of gene prioritization is similar to supervised learning: candidate genes are ranked with respect to training genes based on a dissimilarity index. In the context of this study, training genes correspond to the genes of a reconstructed ABC system, and candidate genes correspond to all the ABC coding genes of the genomes investigated (including the training genes).
Each candidate gene is considered in turn and its dissimilarity score is computed as the average of its dissimilarity with each training gene. If the gene considered is part of the training genes then its dissimilarity to itself is not taken into account when computing its dissimilarity score.
Each data source leads to a list of scored genes. To merge them, we propose to perform a linear discriminant analysis (LDA). Indeed, each data source can be seen as a dimension and the similarity scores of a gene as its coordinates in this multidimensional space. We then face a standard classification task with two classes: genes belonging to the reconstructed system or genes belonging to another system. Thus, we are looking for the combination of the scores obtained on each data source that will best separate the training gene from the others. By performing an LDA, data points are projected in a plane that has C-1 axes, C being the number of classes. Thus, we obtain a one dimensional projection of the candidate genes corresponding to the weighted fusion of the previously performed prioritizations. This has the advantage of providing both the contribution of each data source, and the fusion of the scores by applying the transformation (see Figure 1 for an illustration).
If the system was well assembled, the top genes are expected to be the training genes. Otherwise, the result might point to errors in the reconstruction.

Genome-wide ABC Systems Semi-automatic Validation

In the previous section, we described how prioritization could be used to evaluate the reassembled systems. Either all the genes of a reassembled system rank at the top of the fusion in which case the system can be assumed to be well reconstructed, or there can be other genes that ranked better. In the latter case, this means that the data sources suggest that other genes are more probably associated to part of the system. However, these other genes were reassembled into other systems that might or might not be also misassembled. Thus, these other systems should be considered in parallel. Once all the systems of the studied genome have been prioritized, the outcome of the validations can be summarized as a directed graph modeling the possible reconstruction errors. In this graph, nodes correspond to systems and edges reflect the fact that other genes ranked better than the training genes. In other words, an edge from system *A* to system *B* is added if genes involved in *B* ranked better than some genes involved in system *A* during the evaluation of *A*. Such a representation allows to apprehend the validation results all at once.
Three cases can occur. First, a node is isolated meaning that the corresponding system was well reconstructed. Second, there is an edge from a node to another (A → B). In this case, we can consider that B is well reconstructed and its genes cannot be involved in A. As a result, A is also correctly assembled. Third, there can be cycles (A → B → C → A or A → B → A) that reveal more complex situations that require expert manual intervention.

Application to ABC Systems

To illustrate the proposed strategy, we evaluated the 43 expertized ABC systems of at least two genes from *E. coli* K12 MG1655. The graph based summary of the evaluation is given in Figure 2: our strategy identifies 37 well reconstructed systems (18 isolated nodes and 19 nodes not within cycles), and 6 systems that would have required manual inspection (nodes within cycles). The systems involved in the same cycles appear to belong to the same subfamily of transporters. The fepC system is characterized as an iron (Fe^{3+})-enterobactin transporter and fhuC is characterized as an iron (Fe^{3+})-hydroxamate transporter. glnQ and artP are both transporting amino-acids: glutamine for glnQ and arginine for artP. ugpC transports glycerol-phosphate while the compound transported by ycjV is unknown.

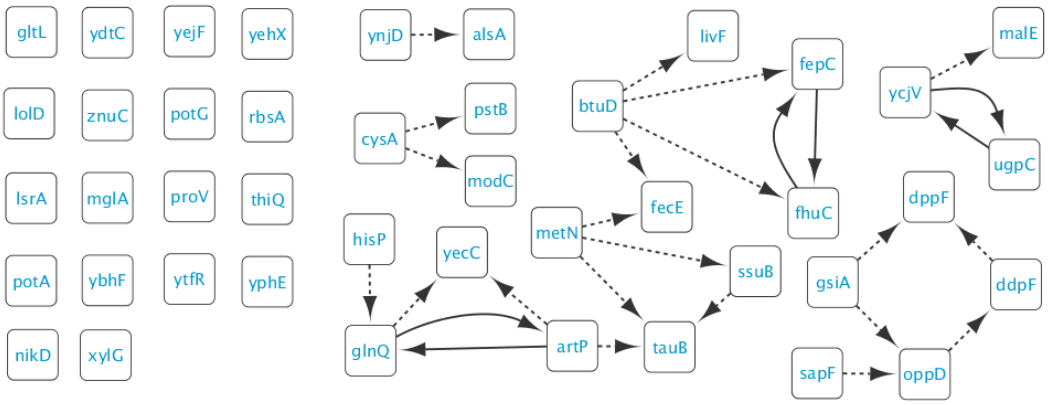


Figure 2. Graph based summary of *E. coli* ABC systems reconstruction evaluation. Nodes correspond to evaluated systems. Edges indicate that genes of other systems ranked better than genes of the reassembled system. Edges with dashed line are not part of a cycle.

Discussion and Conclusions

We presented a strategy to automate the validation of reconstructed systems or point the expert to possible errors. It is inspired by gene prioritization. The main novelties are an automatic weighing of the contribution of each data source based on a linear discriminant analysis, and a post-processing step for the genome-wide validation of the reconstruction based on a graph analysis.

The weighing obtained by the discriminant analysis is similar to the kernel combination described by De Bie and colleagues[14] while having the advantage to be simpler to implement. Compared to Endeavour[3], our approach benefits from the fact that all the systems of a genome are considered at once which allows to resolve conflicts in the prioritized list obtained by a graph analysis.

While the strategy described has been developed for the validation of reconstructed ABC systems, it is generic and could be applied to other types of systems.

Acknowledgments

This work was supported by the Centre National de la Recherche Scientifique (CNRS) and a grant from the Université Paul Sabatier – Toulouse III.

Bibliography

1. Fichant, G., Basse, M.-J. & Quentin, Y. ABCdb: an online resource for ABC transporter repertories from sequenced archaeal and bacterial genomes. *FEMS Microbiol. Lett.* **256,** 333–339 (2006).
2. Pagani, I. *et al.* The Genomes OnLine Database (GOLD) v.4: status of genomic and metagenomic projects and their associated metadata. *Nucleic Acids Res.* **40,** D571–579 (2012).
3. Aerts, S. *et al.* Gene prioritization through genomic data fusion. **24,** 537–544 (2006).
4. Barrett, T. *et al.* NCBI GEO: archive for functional genomics data sets—update. *Nucleic Acids Res.* **41,** D991–D995 (2013).
5. Gentleman, R. C. *et al.* Bioconductor: open software development for computational biology and bioinformatics. *Genome Biol.* **5,** R80 (2004).
6. Gautier, L., Cope, L., Bolstad, B. M. & Irizarry, R. A. affy—analysis of Affymetrix GeneChip data at the probe level. *Bioinformatics* **20,** 307–315 (2004).
7. Oti, M., van Reeuwijk, J., Huynen, M. & Brunner, H. Conserved co-expression for candidate disease gene prioritization. **9,** 208 (2008).
8. The. Gene Ontology: tool for the unification of biology. **25,** 25–29 (2000).
9. Camon, E. *et al.* The Gene Ontology Annotation (GOA) Database: sharing knowledge in Uniprot with Gene Ontology. *Nucleic Acids Res.* **32,** D262–D266 (2004).
10. Guzzi, P. H., Mina, M., Guerra, C. & Cannataro, M. Semantic similarity analysis of protein data: assessment with biological features and issues. *Brief. Bioinform.* **13,** 569–585 (2012).
11. Jensen, L. *et al.* STRING 8--a global view on proteins and their functional interactions in 630 organisms. *Nucleic Acids Res* **37,** D412–D416 (2009).
12. Ferrer, L., Dale, J. M. & Karp, P. D. A systematic study of genome context methods: calibration, normalization and combination. *BMC Bioinformatics* **11,** 493 (2010).
13. Calderon, V., Barriot, R., Quentin, Y. & Fichant, G. Crossing Isorthology and Microsynteny to Resolve Multigenic Families Functional Annotation. in *2012 23rd Int. Workshop Database Expert Syst. Appl.* **0,** 440–444 (IEEE Computer Society, 2011).
14. De Bie, T., Tranchevent, L.-C., van Oeffelen, L. M. M. & Moreau, Y. Kernel-based data fusion for gene prioritization. *Bioinforma. Oxf. Engl.* **23,** i125–132 (2007).

Metabolic Meta-Reconstruction and Community Modeling of Intestinal Microbes

Eugen Bauer and Ines Thiele
Luxembourg Centre for Systems Biomedicine,
University of Luxembourg, Luxembourg
{eugen.bauer,ines.thiele}@uni.lu

The human intestine harbors a highly diverse microbial community, which is, in cell number, ten times larger than the eucaryotic cells in the human body [4]. Naturally these highly abundant internal microbes play important roles in human health and disease. The main functional roles of the gut microbiota can be divided into three categories: metabolic, protective and regulatory functions [6].

In terms of their metabolic importance intestinal microbes provide the host with the ability to extract energy of otherwise indigestable polysaccharides [11]. Bacteria degrade these substrates into short chain fatty acids (SCFA), which can be taken up by the intestinal epithelium for energy conversion [5]. Another benefit of the SCFAs is the acidification of the gut lumen, which inhibits the growth of certain pathogens [9]. In addition to this indirect protective function the growth of pathogens can also be directly inhibited with the bacterial production of antimicrobial peptides (AMP), such as bacteriocins [10]. AMPs can also be produced by the regulatory innate immune response of the host, which development is dependent on the microbial gut community and vice versa [3]. Another example of this cross-talk between host and symbionts can be found in the gut-brain axis. Here, the gut microbiota plays a pivotal role in the stress regulation of the central nervous system [7].

The manifold host-microbiota interactions are tightly regulated by the host as well as the bacterial community itself. Any perturbation of this host-symbiont equilibrium caused by drugs, nutritional or genetic changes can result in various human disease, such as obesity, chronic diarrhea, and inflammatory bowel disease [8]. Studies like the human microbiome project [17] or MetaHit [18] enhanced our understanding of the microbial ecology in these diseases while constantly providing reference genomes of selected organisms. However, our mechanisitic insight into community structure and the underlying metabolic interactions is still incomplete. In this context, recent advances in constrained based reconstruction and analysis (COBRA) of genome scale metabolic models [2] have been applied to study host-microbe interactions and predicting phenotypic changes [1].

In this study we automatically reconstructed genome scale metabolic models based on the genomes of 301 representative intestinal microbes [13] using the pipeline SEED [12]. We augmented the metabolic models with manual curation and gap filling of essential reactions allowing growth under anaerobic conditions for those microbes, which were predicted to survive under anoxic conditions. The curated metabolic models were then subjected to a linear optimization of the biomass objective function using flux balance analysis [14]. The growth conditions were further adjusted to reasonable conditions by changing the bounds of exchange reactions.

To assess individual differences within the various metabolic models we used a classification based on unsupervised machine learning methods of the reaction and metabolite content. Here, we computed the Jaccard Index of all pair-wise comparisons between the microbes based on the presence and absence of the overall metabolites and reactions [15]. The resuling distance matrix was then analysed further using a linear approach with principle coordinate analysis as well as a non-linear approach with tSNE [16].

Additionally, we subjected each model to varying *in silico* growth conditions to assess potential auxotrophies and prototrophies on certain metabolites and compare the predictions with the reaction/metabolite

Figure 1: Overview of the automatically reconstructed genome scale metabolic models of 301 intestinal microbes and their taxonomic placement. The annotation rings represent the relative genome size, number of metabolic genes, reaction number and *in silico* growth rate. Known pathogenic and probiotic species are colored in red and green respectively.

classification. We further investigated the different classification schemes (reaction, metabolite, growth conditions) by a comparative hierarchical clustering. To elucidate cooperative microbial interactions we combined all 301 metabolic models to one joint model with an exchange compartment dedicated for metabolic interactions. As a negative control we created a model without allowing these interactions. Based on the metabolic fluxes in the interaction inducing compartment we constructed a metabolic interaction network representing the metabolic exchange across the different microbes. The nodes in this network comprise the different microbial species, while the edges represent the exchanged metabolites. According to this network, we studied basic topological features to identify important community members.

Overall, the metabolic models showed siginificant differences in their genome size, gene content, reaction number and *in silico* growth rate (Figure 1). The differences observed in the reaction and metabolite content recapitulated the taxonomic placement and phenotype of the individual microbes and might serve as a classification scheme to characterize newly sequenced microbes by their reaction content. In particular, we found non-linear machine learning techniques to produce better results, than conventional methods, indicating the non-linearity of the underlying problem. Furthermore, the differences observed in the usage of particular amino acids and carbohydrates also agreed with the taxonomic placement and highlight taxa specific auxotrophies. Furthermore, this results indicate a strong phylogentic signal within genomic sequences, which substantiate the usability of metabolic reconstructions based on phylogeny [19].

While fundamentally different in their metabolic capacities, microbes of different taxa were found to exhibit more metabolic interactions with each other, compared to taxonomically more similar partners. These findings are in concordance with previous studies [15] and indicate a devision of labour within the community, which is distributed throughout different taxa. This demonstrates the need of microbial consortia, in particular within the human gut, to cotain a certain taxonomic diversity of community members to ensure a metabolic flexibity.

This study presents the first metabolic reconstruction and joint simulation of 301 individual genome scale metabolic models. By combining methods of unsupervised machine learning and constrained based modeling we were able to elucidate the manifold metabolic interactions within intestinal microbial communities. The results obtained in our study are relevant for potential disease treatments as well as for biotechnology applications.

References

[1] Thiele, I., Heinken, A., and Fleming, R. M. (2013). A systems biology approach to studying the role of microbes in human health. Current opinion in biotechnology, 24(1), 4-12.

[2] Schellenberger, J., Que, R., Fleming, R. M., Thiele, I., Orth, J. D., Feist, A. M., et al (2011). Quantitative prediction of cellular metabolism with constraint-based models: the COBRA Toolbox v2. 0. Nature protocols, 6(9), 1290-1307.

[3] Round, J. L., Lee, S. M., Li, J., Tran, G., Jabri, B., Chatila, T. A., and Mazmanian, S. K. (2011). The Toll-like receptor 2 pathway establishes colonization by a commensal of the human microbiota. Science, 332(6032), 974-977.

[4] Bengmark, S. (1998). Ecological control of the gastrointestinal tract. The role of probiotic flora. Gut, 42(1), 2-7.

[5] Martens, E. C., Lowe, E. C., Chiang, H., Pudlo, N. A., Wu, M., McNulty, N. P., et al (2011). Recognition and degradation of plant cell wall polysaccharides by two human gut symbionts. PLoS biology, 9(12), e1001221.

[6] Guarner, F., and Malagelada, J. R. (2003). Gut flora in health and disease. The Lancet, 361(9356), 512-519.

[7] Foster, J. A., and McVey Neufeld, K. A. (2013). Gut-brain axis: how the microbiome influences anxiety and depression. Trends in neurosciences, 36(5), 305-312.

[8] Kamada, N., Seo, S. U., Chen, G. Y., and Nunez, G. (2013). Role of the gut microbiota in immunity and inflammatory disease. Nature Reviews Immunology, 13(5), 321-335.

[9] Robles Alonso, V., and Guarner, F. (2013). Linking the gut microbiota to human health. British Journal of Nutrition, 109(S2), S21-S26.

[10] Aziz, Q., Dore, J., Emmanuel, A., Guarner, F., and Quigley, E. M. M. (2013). Gut microbiota and gastrointestinal health: current concepts and future directions. Neurogastroenterology & Motility, 25(1), 4-15.

[11] Blaut, M., and Clavel, T. (2007). Metabolic diversity of the intestinal microbiota: implications for health and disease. The Journal of nutrition, 137(3), 751S-755S.

[12] Devoid, S., Overbeek, R., DeJongh, M., Vonstein, V., Best, A. A., and Henry, C. (2013). Automated genome annotation and metabolic model reconstruction in the SEED and Model SEED. In Systems Metabolic Engineering (pp. 17-45). Humana Press.

[13] Qin, J., Li, Y., Cai, Z., Li, S., Zhu, J., Zhang, F., et al (2012). A metagenome-wide association study of gut microbiota in type 2 diabetes. Nature, 490(7418), 55-60.

[14] Orth, J. D., Thiele, I., and Palsson, B. O. (2010). What is flux balance analysis?. Nature biotechnology, 28(3), 245-248.

[15] Mazumdar, V., Amar, S., and Segre, D. (2013). Metabolic proximity in the order of colonization of a microbial community. PloS one, 8(10), e77617.

[16] Van der Maaten, L., ans Hinton, G. (2008). Visualizing data using t-SNE. Journal of Machine Learning Research, 9(2579-2605), 85.

[17] Turnbaugh, P. J., Ley, R. E., Hamady, M., Fraser-Liggett, C., Knight, R., and Gordon, J. I. (2007). The human microbiome project: exploring the microbial part of ourselves in a changing world. Nature, 449(7164), 804.

[18] Ehrlich, S. D. (2011). MetaHIT: The European Union Project on metagenomics of the human intestinal tract. In Metagenomics of the Human Body (pp. 307-316). Springer New York.

[19] Pitkaenen, E., Jouhten, P., Hou, J., Syed, M. F., Blomberg, P., Kludas, J., et al (2014). Comparative Genome-Scale Reconstruction of Gapless Metabolic Networks for Present and Ancestral Species. PLoS computational biology, 10(2), e1003465.

Meta-analysis of Genome-Wide Association Studies with Multivariate Traits

Anna Cichonska[1,2*], Pekka Marttinen[2], Samuli Ripatti[1,3,4], Juho Rousu[2], Matti Pirinen[1]

[1] Institute for Molecular Medicine Finland FIMM, University of Helsinki, Finland

[2] Helsinki Institute for Information Technology HIIT, Department of Information and Computer Science, Aalto University, Finland

[3] Hjelt Institute, University of Helsinki, Finland

[4]Department of Human Genetics, Wellcome Trust Sanger Institute, Wellcome Trust Genome Campus, United Kingdom

*To whom correspondence should be addressed; E-mail: anna.cichonska@helsinki.fi

Introduction

Most human diseases have a strong genetic component. The aim of Genome-Wide Association Studies (GWAS) is to find genetic variations correlated with a particular trait. A common approach is to use univariate phenotypes (e.g. a single metabolite measurement or a binary disease indicator) and perform simple univariate tests between genotype-phenotype pairs. There are many summary-level results of such analyses publicly available. However, rapidly developing technologies provide us with a growing number of phenotypic features, including serum metabolomic profiles [1] or measurements in 3D space, representing for instance gray matter intensities in the brain [2]. Currently, there is an interest to investigate associations using multi-dimensional phenotypes. It has been shown that utilising multivariate phenotype representation can result in increased power and richer findings in the association analysis [1]. However, low sample sizes in individual studies and public unavailability of complete multivariate data are the key limitations. Meta-analysis would overcome the problem of small sample size if only it could be carried out using univariate summary statistics that are publicly available. Moreover, meta-analysis offers benefits analogous to utilising multi-dimensional phenotypes [3].

The goal of this work was to establish a computational approach for multivariate meta-analysis of GWA studies, based on available summary-level results of univariate analysis.

Methods

Let $X^{(i)}_{N\times G}$ and $Y^{(i)}_{N\times P}$ denote the genotype and phenotype matrices storing the original data coming from i=1,...,m studies conducted on the same topic, $N^{(i)}$ the number of samples in the ith study, G and P the number of genotypic and phenotypic variables, respectively. Genotype is typically represented as Single Nucleotide Polymorphisms (SNPs). $X^{(i)}$ and $Y^{(i)}$ are standardised, such that the mean of each variable is 0 and the standard deviation is 1.

Univariate analysis is performed by using a simple linear regression, testing for a presence of the relationship between each pair of SNP $\boldsymbol{x}_g^{(i)}$ and phenotype $\boldsymbol{y}_p^{(i)}$ separately. In order to test for an association, the following model is used:

$$\boldsymbol{y}_p^{(i)} = \beta_0 + \beta_1 \boldsymbol{x}_g^{(i)} + \epsilon.$$

Coefficient β_0 is an intercept on the y-axis, and β_1, corresponding to the slope of the regression line, depicts the size of the effect. ϵ is an error term (noise). Coeffcients β_1 for all possible genotype-phenotype pairs can be stored in the matrix form:

$$B^{(i)} = \frac{\mathrm{Cov}(X^{(i)}, Y^{(i)})}{\mathrm{Cov}(X^{(i)}, X^{(i)})} = \frac{\mathrm{Cov}(X^{(i)}, Y^{(i)})}{1} = \\ = \frac{X^{(i)T} Y^{(i)}}{N^{(i)} - 1} = \Sigma^{(i)}_{XY}.$$

$\Sigma^{(i)}_{XY}$ denotes the cross-covariance matrix. 1 is subtracted from $N^{(i)}$ in order to get an unbiased estimator of the sample covariance.

In case of N observations in the sample, there are $N-1$ degrees of freedom.

Often, original data $X^{(i)}$ and $Y^{(i)}$ are publicly unavailable, and low sample size does not allow the researchers to take an advantage of the multivariate analysis of a single study that they have an access to. Thus, our approach is based on cross-covariance matrices, $\Sigma_{XY}^{(i)}$, storing available summary-level results of univariate analysis. In this part of the survey, we assume that we also have genotypic and phenotypic correlation structures $\Sigma_{XX}^{(i)}$ and $\Sigma_{YY}^{(i)}$, and that original data matrices, $X^{(i)}$ and $Y^{(i)}$, are unknown.

Our approach uses canonical correlation analysis (CCA) with a novelty that it operates on pooled covariance matrices, C_{XY}, C_{XX}, C_{YY}, rather than requiring all of the original genotype $X^{(i)}$ and phenotype $Y^{(i)}$ data. Covariance matrices of the same type, coming from m studies, are pooled using a weighted average:

$$C_{XY} = \frac{(N^{(1)}-1)\Sigma_{XY}^{(1)} + ... + (N^{(m)}-1)\Sigma_{XY}^{(m)}}{N_t - m},$$

where $N_t = N^{(i)} + ... + N^{(m)}$. The formula for C_{XX} and C_{YY} is analogical to the one above. Weighted average is used in order to account for the standard error of the covariance estimate (the lower $N^{(i)}$, the higher the error).

CCA is a multivariate technique designed for analysing paired datasets by detecting associations between two groups of variables X and Y, where X and Y constitute two different views of the same object [4]. The aim is to find a linear combination of columns of each matrix, which corresponds to finding vectors $\boldsymbol{a} \in \mathbb{R}^G$ and $\boldsymbol{b} \in \mathbb{R}^P$ in the data space, such that the canonical correlation ρ between X and Y is maximised:

$$\rho_l = \frac{(X\boldsymbol{a}_l)(Y\boldsymbol{b}_l)}{\|X\boldsymbol{a}_l\| \, \|Y\boldsymbol{b}_l\|},$$

where l=1,...,k. k=min$\{G, P\}$, what in practice means that the number of canonical projections that can be found by CCA is equal to the smallest of the ranks of matrices X and Y. Vectors $\boldsymbol{a}_l$ and $\boldsymbol{b}_l$ are called canonical weights. Magnitudes of the elements stored in these vectors can be used to identify variables that provide large contribution to the canonical correlation. In our approach, pooled covariance matrices are first used to compute matrix K:

$$K = C_{XX}^{-1/2} C_{XY} C_{YY}^{-1/2}.$$

Next, Singular Value Decomposition (SVD) theorem is applied. It allows to decompose K into three matrices:

$$K = (\boldsymbol{\alpha}_1, ..., \boldsymbol{\alpha}_k) D (\boldsymbol{\gamma}_1, ..., \boldsymbol{\gamma}_k)^T.$$

Above, $\boldsymbol{\alpha}_l$ and $\boldsymbol{\gamma}_l$ are the standardised eigenvectors of KK^T and K^TK, respectively. D is a diagonal matrix of square roots of the corresponding eigenvalues $\sqrt{\lambda_1}, ..., \sqrt{\lambda_k}$, and ρ_l=$\sqrt{\lambda_l}$. Canonical weights corresponding to the lth projection are computed based on the eigenvectors $\boldsymbol{\alpha}_l$ and $\boldsymbol{\gamma}_l$ [5]:

$$\boldsymbol{a}_l = C_{XX}^{-1/2} \boldsymbol{\alpha}_l,$$
$$\boldsymbol{b}_l = C_{YY}^{-1/2} \boldsymbol{\gamma}_l.$$

The canonical correlation ρ_l is defined as:

$$\rho_l = \frac{\boldsymbol{a}_l^T C_{XY} \boldsymbol{b}_l}{\sqrt{\boldsymbol{a}_l^T C_{XX} \boldsymbol{a}_l} \sqrt{\boldsymbol{b}_l^T C_{YY} \boldsymbol{b}_l}}.$$

Correlation ρ is maximised by selecting $\boldsymbol{a}_1$ and $\boldsymbol{b}_1$. Subsequent canonical projections are uncorrelated with each other.

We focused on two types of analyses:

- one SNP tested for an association with the set of phenotypic variables (analysis based on matrices C_{XY} and C_{YY});
- the set of SNPs tested for an association with the set of phenotypic variables (C_{XY}, C_{XX}, C_{YY}).

The goal is to find SNPs correlated with a particular trait across studies.

Statistical significance. In order to investigate which of the obtained canonical correlations are statistically significant, we used Bartlett's χ^2 test [6]. It is performed in two steps. First, the test's statistic Wilk's Lambda is computed based on the eigenvalues λ_l:

$$\Lambda_l = \prod_l^k (1 - \lambda_l).$$

Next, Bartlett's χ^2 approximation to Wilk's Lambda is calculated:

$$\chi_l^2 = -\Big[(N_t - 1) - \frac{G + P + 1}{2}\Big] \ln \Lambda_l.$$

Wilk's Lambda is distributed approximately as χ^2 with $(G - l + 1) \times (P - l + 1)$ degrees of freedom. The null hypothesis related to ρ_l states that subsequent $(k - l + 1)$ canonical correlations are equal to 0, meaning that two types of variables in the paired data set are not related. Each canonical correlation ρ_l is tested by the same test statistic, using its corresponding eigenvalue λ_l. However, it is a sequential process where the contribution from the previous $l - 1$ canonical projections is removed before χ_l^2 is computed. That is why the number of degrees of freedom is also being reduced successively. χ_1^2 corresponds to the test that all canonical correlations are equal to 0. χ_2^2 corresponds to the test that consecutive correlations, starting from the second one, are equal to 0, and so forth.

Data sets. In order to test our approach, we used data coming from three studies with increasing sample sizes ($\mathrm{N}^{(1)} = 2390$, $N^{(2)} = 3661$, $N^{(3)} = 4702$). We removed from the data SNPs with minor allele frequencies lower than 5%. Phenotypes are represented as metabolites' levels. In total there are 259 SNPs (corresponding to one gene) and 81 phenotypic variables common for three studies. Such setup corresponds to many actual population-based cohorts with hundreds of phenotypes measured. For each study, we have matrices $\Sigma_{XY}^{(i)}$, $\Sigma_{XX}^{(i)}$ and $\Sigma_{YY}^{(i)}$.

Results

One SNP vs. the set of phenotypic variables. First, each SNP was tested for an association with the set of 81 phenotypic variables. In comparison to analyzing single survey separately, more significant associations were identified by doing the meta-analysis of three studies (Table 1). Moreover, we performed the meta-analysis by applying CCA on a concatenation of the samples (pooling together individuals from three studies), assuming that original data $X^{(i)}$ and $Y^{(i)}$ are known. The result was the same as in case of using our approach.

study_1	study_2	study_3	meta-analysis
4	11	12	**37**

Table 1: Number of SNPs significantly associated with the set of phenotypic variables ($P < 5 \times 10^{-8}$). The significance level is lower than the standard one of 5×10^{-2} because any SNP at random from the genome has the same probability of being associated with the phenotype, and there are approximately 10^6 uncorrelated common SNPs in the human genome.

The set of SNPs vs. the set of phenotypic variables. Next, the set of SNPs was tested for an association with the set of 81 phenotypic variables (meta-analysis of three studies). Canonical correlation analysis identified three significant canonical projections ($P < 5 \times 10^{-2}$). Associations found by the first two projections are shown in Figure 1. We implemented associations visualisation based on [7] (correlation circle plots). It allows to present the associations found by two selected projections (dimensions), l_1, l_2, in one plot. The idea is to calculate the correlation between each original variable, $\boldsymbol{x}_g$, $\boldsymbol{y}_p$, and its associated canonical variable, $\boldsymbol{u}_l$, $\boldsymbol{v}_l$, respectively, where $\boldsymbol{u}_l = X\boldsymbol{a}_l$, $\boldsymbol{v}_l = Y\boldsymbol{b}_l$. However, here X and Y are unknown, so our implementation is based on the *covariance space*. Each point in the plot represents one variable. Coordinates of jth genotypic variable in the plot are calculated as $(\mathrm{corr}(\boldsymbol{c}_{XX}^j, C_{XX}\boldsymbol{a}_{l_1}), \mathrm{corr}(\boldsymbol{c}_{XX}^j, C_{XX}\boldsymbol{a}_{l_2}))$. Coordinates of the jth phenotypic variables are computed analogically: $(\mathrm{corr}(\boldsymbol{c}_{YY}^j, C_{YY}\boldsymbol{b}_{l_1}), \mathrm{corr}(\boldsymbol{c}_{YY}^j, C_{YY}\boldsymbol{b}_{l_2}))$. Each point can be thought as a vector that starts in the origin of the plot. Correlation between variables is approximated by the inner product between their associated vectors, e.g. the angle between green and blue vectors in Figure 1 is sharp, indicating strong positive correlation between two SNPs (212, 213) and a metabolite marked with *M.HDL.PL*.

Ongoing Work

Currently, we are working on the method extension, where sample covariance matrices are estimated ($\hat{C}_{XX}$, $\hat{C}_{YY}$), and thus not needed for individual studies. The goal is to perform the analysis only based on the univariate results of separate studies, $\Sigma_{XY}^{(i)}$, and two estimated matrices: $\hat{C}_{XX}$ and $\hat{C}_{YY}$. It is known that SNPs correlation structure is the same for a given pop-

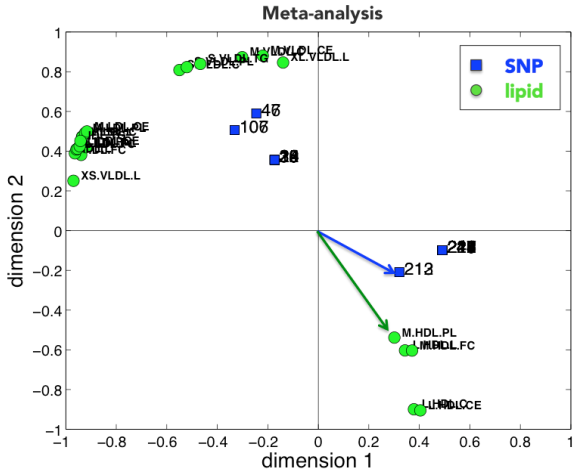


Figure 1: SNPs-phenotypes association profile's visualization. There are many SNPs on top of each other, meaning that they are very strongly correlated due to the Linkage Disequilibrium (LD).

ulation. Thus, $\hat{C}_{XX}$ can be estimated from a reference database representing the study population, such as the *1000 Genomes* database (www.1000genomes.org). $\hat{C}_{YY}$ can be computed based on C_{XY}, $\hat{C}_{YY}$=$C_{XY}^T C_{XY}$, where columns of C_{XY} are vectors of unit length. The higher the number of genotypic variables, the lower the error of the estimate. Hence, $\hat{C}_{YY}$ can be computed based on the full SNP data, even if only a subset of SNPs at a time is taken into the analysis. However, the problem is computationally non-trivial because matrices $\hat{C}_{XX}$ and $\hat{C}_{YY}$ cannot be just plugged into the analysis. First, we need to find the nearest positive semi-definite full covariance matrix and reduce the condition number of the matrix K (which increases when using $\hat{C}_{XX}$ and $\hat{C}_{YY}$ instead of actual C_{XX} and C_{YY}). Condition number can be used e.g. to characterise matrix stability and numerical feasibility of factorisation algorithms, like SVD. We are working with covariance shrinkage methods and semi-definite programming (SDP). The analysis is based on simulated and real data with different number of samples and variables.

Figure 2 shows the significant reduction of the data in our approach.

Discussion

We introduced a computational approach for multivariate meta-analysis of GWA studies, based on the available results of univariate analysis. Our

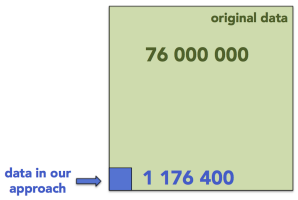


Figure 2: Example of data reduction in our approach: meta-analysis of 45 studies (200000 individuals), 300 SNPs - 1 gene (X), 80 metabolite levels (Y). Original data: 200000×(300+80)=76e6; data in our approach, taking into the account $\Sigma_{XY}^{(i)}$ (for each study), $\hat{C}_{XX}$, $\hat{C}_{YY}$: 45×300×80+300×300+80×80≈1.2e6.

approach has an advantage of sharing covariance matrices instead of the full data. It negates the problem of public unavailability of the original data, and has a benefit of using less data than standard multivariate approaches that are based on the original data. Moreover, meta-analysis allows to tackle the problem of low sample sizes in individuals studies. In case of GWAS, both the meta-analysis and multidimensional phenotypes are of high importance as they can lead to richer findings, which can be further used for instance in designing effective therapies. Recent studies have shown that it is possible to use GWAS to identify novel drug targets [8].

References

[1] Inouye M., et al.: *Novel Loci for metabolic networks and multi-tissue expression studies reveal genes for atherosclerosis.* PLoS Genetics 8.8, e1002907 (2012).

[2] Wang Y., et al.: *Random forests on Hadoop for Genome-Wide Association Studies of multivariate neuroimaging phenotypes.* BMC Bioinformatics 14.16, 1-15 (2013).

[3] Global Lipids Genetics Consortium: *Discovery and refinement of loci associated with lipid levels.* Nature Genetics (2013).

[4] Hardoon D. R., et al.: *Canonical correlation analysis: an overview with application to learning methods.* Neural Computation 16.12, 2639-2664 (2004).

[5] Marttinen P., et al.: *Genome-wide association studies with high-dimensional phenotypes.* Statistical Applications in Genetics and Molecular Biology 12.4, 413-431 (2013).

[6] Lee H. B., et al.: *The application of canonical correlation to two-dimensional contingency tables.* Tutorials in Quantitative Methods for Psychology 6.1, 1-15 (2010).

[7] Gonzalez I., et al.: *Visualising associations between paired omics data sets.* BioData Mining 5.1, 1-23 (2012).

[8] Okada Y., et al.: *Genetics of rheumatoid arthritis contributes to biology and drug discovery.* Nature 506, 376-381 (2013).

Computationally Efficient Test for Gene Set Dysregulation

Adrien Dessy and Pierre Dupont

Université catholique de Louvain - ICTEAM/Machine Learning Group
Place Sainte Barbe 2 bte L5.02.01, B-1348 Louvain-la-Neuve, Belgium
adrien.dessy@uclouvain.be

Abstract. The identification of genetic regulatory pathways whose structure differs across biological conditions provides significant insights about organisms and diseases functioning at molecular level. In this paper, we propose a computationally efficient test to assess from gene expression data if a given group of genes is differentially regulated between two conditions. The method yields promising results in terms of precision and recall on real datasets.

1 Introduction

Differential analysis of Gene Regulatory Networks (GRNs) has been raising a growing interest lately. There is not yet a standard definition to this problem, but the high-level shared goal is to assess if the interactions or associations of genes differ between two or more biological conditions. This type of analysis can be performed at the network level, for a given group of genes (subnetwork) or for a specific interaction between two genes for instance.

Most of the proposed methods have in common that they compare networks infered for each condition from gene expression data [1,2]. Hence, they rely on network inference techniques or association measures between genes. The comparison is then based on a differentiation score whose significance is assessed by a permutation test. The first contribution of this work is to propose an alternative permutation test that is more computationally efficient.

The definition of differential network analysis slightly differs between studies. Either because they operate at different network levels, or because some studies test a given component for differentiation, while others try to discover differentiated parts of the network. Gill et al. [1] proposes three statistical tests to assess whether the modular structures of two networks are different, whether the connectivity of a group of genes or the connectivity of a single gene has changed. Liu et al. [3] describes a procedure to determine if genes of a pathway are differentially wired between two conditions. If so, a differential network is build by testing dysregulation of each interaction in the pathway using a t-test. The DINA procedure proposed by Gambardella et al. [2] also intends to assess whether co-regulation among a given set of genes depends on the condition, but accross multiple networks. Amar et al. [4] presents an algorithm to extract differential gene clusters. Our work is closely related to these techniques, especially [1,2]. In this paper, we describe a method to :

(a) test if a given group of genes (a module) is differentially regulated between two conditions;
(b) rank modules by observed dysregulation level.

In Section 2, we describe the details of the method. Section 3 discusses the results of our experiments on two real gene expression datasets. Finally, Section 4 presents the conclusions of this work and suggests some future works.

2 Method

The method that we propose is summarized in Figure 1. It consists of three main steps. Firstly, GRNs are infered from gene expression data for both conditions. Subsequently, a differentiation score is computed by comparing the GRNs. The significance of this score is finally estimated through a permutation test.

2.1 Inference of gene regulatory networks

For each biological condition, a GRN is infered using the MRNET approach [5]. MRNET inference consists in performing a sequence of mRMR gene selection procedures with each gene as output variable. mRMR algorithm selects iteratively variables depending on the previously selected variables (gene expression profiles in our case). At each iteration, it selects the gene that maximizes an objective function measuring a trade-off between the mutual information with the target gene (relevance) and the mean mutual information with the already selected genes (redundancy).

For the sake of computational efficiency, we made assumption of data normality. Under this assumption, mutual information can readily be computed as

$$\mathbf{MI}_{ij} = -\frac{1}{2}\ln(1-\rho_{ij}^2)$$

where ρ_{ij} is the pearson correlation between genes i and j.

This step produces two adjacency matrices $\mathbf{A}^1$ and $\mathbf{A}^2$ representing both GRNs. These matrices are symetric with null diagonal and their entries are in the range $[0, 1]$. Note that MRNET is quite an arbitrary choice. Another GRN inference method could have been chosen.

2.2 Differentiation score

The differentation score, denoted by s_Δ, represents the differentiation level observed between GRNs with respect to a module. A large score s_Δ indicates an important differentiation. In mathematical terms, it is computed as three-argument function :

$$s_\Delta = f_\Delta(\mathbf{A}^1, \mathbf{A}^2, \mathcal{M})$$

where $\mathcal{M}$ is the group of genes of interest.

The computation of s_Δ can be decomposed into two steps. The first part is applied separately to each network and aims to extract statistics that depend on the topology of the module. The second part compares these statistics to produce the differentiation score.

In the first step, network statistics are computed as $\mathbf{s}^1 = f(\mathbf{A}^1, \mathcal{M})$ and $\mathbf{s}^2 = f(\mathbf{A}^2, \mathcal{M})$, such that $\mathbf{s}^1$ and $\mathbf{s}^2$ are real vectors of same length ($\mathbb{R}^K$) for any instantiation of the scoring function f. The scores are then combined as

$$\begin{aligned} s_\Delta &= f_\Delta(\mathbf{A}^1, \mathbf{A}^2, \mathcal{M}) \\ &= \sum_{k=1}^{K} |\mathbf{s}^1_k - \mathbf{s}^2_k| = ||\mathbf{s}^1 - \mathbf{s}^2||_1 \end{aligned}$$

where $||.||_1$ denotes the $\ell 1$ norm.

Instead of this two-stage scheme, we could have adapted graph kernels to compute the differentiation scores. But, for the sake of computational efficiency, we chose this simple approach as a first step.

We now introduce in the rest of the section different instantiations of the scoring functions f. Several variants have been explored, but for the sake of brevity only simple graph statistics based on node degree will be reported here.

Degree The function f_{degree} returns the degree of each node in $\mathcal{M}$. More precisely, let's define $\mathbf{s} = f_{degree}(\mathbf{A}, \mathcal{M})$. Without loss of generality, we can reorder genes such that $\mathcal{M} = \{1, 2, \ldots, M\}$. We have that $\mathbf{s} \in \mathbb{R}^M$ and

$$\mathbf{s}_i = \sum_{j \in \mathcal{M}, i \neq j} \mathbf{A}_{ij}, \quad \forall i \in \mathcal{M}.$$

Notice that this definition uses only weights of edges between genes in $\mathcal{M}$. Furthermore, the score vector $\mathbf{s}$ can be normalised as $\mathbf{s}' = \frac{1}{\max_{i \in \mathcal{M}} \mathbf{s}_i} \mathbf{s}$.

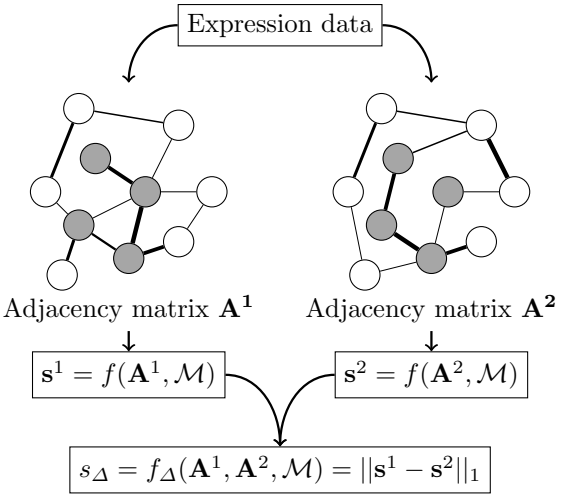


Fig. 1 – Method overview : A GRN is infered for both biological conditions from gene expression data. Given a module $\mathcal{M}$ (grey nodes), network statistics $\mathbf{s}^1$ and $\mathbf{s}^2$ are computed from each GRN and combined to produce the score of differentiation $\mathbf{s}_\Delta$.

Mean degree The function $f_{mean.deg}$ is simply defined as the mean degree of the module, that is

$$s = f_{mean.deg}(\mathbf{A}, \mathcal{M}) = \frac{1}{M} \sum_{m=1}^{M} \mathbf{s}_m^{degree}$$

where $\mathbf{s}^{degree} = f_{degree}(\mathbf{A}, \mathcal{M})$. In this case, the scoring function returns a scalar.

Degree variance The function $f_{deg.var}$ computes the degree variance of the module

$$s = f_{deg.var}(\mathbf{A}, \mathcal{M}) = \frac{1}{M-1} \sum_{m=1}^{M} (\mathbf{s}_m^{degree} - \overline{\mathbf{s}^{degree}})^2$$

where $\overline{\mathbf{s}^{degree}} = f_{mean.deg}(\mathbf{A}, \mathcal{M})$. As the previous scoring function, it returns a scalar.

2.3 Permutation test

A permutation test is performed to assess if the differentiation score s_Δ is significant. The standard approach consists in permuting the class labels N times [1,2,3,4]. This requires to reinfer a pair of GRNs for each permutation. This operation has a complexity of $\Omega(p^2)$ where p is the number of genes and becomes costful for real networks that involve thousands of genes.

Here, we propose an alternative approach that is less computationally expensive. The idea is to sample N random modules $\mathcal{M}_n$ ($\forall n, 1 \leq n \leq N$) of size $|\mathcal{M}|$ from all the available genes. This alternative test postulates that the probability of these random modules being differentiated is very low. Hence, a background distribution of s_Δ can be estimated by computing permutation scores

Kegg ID	Name	Size
hsa04010	MAPK signaling pathway	220
hsa04060	Cytokine-cyt. receptor interaction	217
hsa04110	Cell cycle	117
hsa04115	p53 signaling pathway	37
hsa04151	PI3K-Akt signaling pathway	292
hsa04210	Apoptosis	46
hsa05215	Prostate cancer	59

Table 1 – Kegg pathways used as differentiated modules in prostate cancer compared to healthy condition.

$s^n_\Delta = f_\Delta(\mathbf{A}^1, \mathbf{A}^2, \mathcal{M}_n)$ from these random modules. The test p-value is then defined as $p\text{-}value = \#\{s^n_\Delta \geq s_\Delta\}/N$ where $\#\{s^n_\Delta \geq s_\Delta\}$ is the number of permutation scores greater or equal to the original differentiation score. These p-values can used to rank a set of modules according to their dysregulation level.

3 Experiments

We investigate the performance of our approach on real datasets and compare it with a baseline approach. The experiments show promising results in terms of recall and precision.

3.1 Baseline

In order to validate our approach, we compare it with the following baseline inspired from gene set enrichment analysis [6]. Firstly, we select genes with differentiated expression using a Welch's t-test with Benjamini-Hochberg correction. A hypergeometric test is used to test the significance of the overlap between the selected genes and the genes of a given module. The modules can then be ranked by p-value in ascending order.

3.2 Datasets

We tested our approach on two real gene expression datasets : GSE6919[1] and GSE13159[1], retrieved from InSilico DB [7]. In order to be able to measure precision and recall, a set of differentiated modules as well as a set of undifferentiated modules must be known for each dataset. This information has been retrieved from Kegg database of annotated pathways. A Kegg pathway can be readily converted into a module by considering its set of genes.

GSE6919 : prostate cancer. This dataset is composed of gene expression data from normal and prostate cancer tumor tissues. It consists of 171 samples (18 healthy and 153 cancer) and 8801 genes. For computational reasons, it has been reduced to 2000 genes. The set of differentiated modules was defined from the prostate cancer pathway and its related pathways reported in Table 1. And the set of undifferentiated modules was then formed by selecting randomly 50 other Kegg pathways.

GSE13159 : leukemia. This dataset is part of the MILE Study (Microarray Innovations In LEukemia) program and encompasses 2096 gene expression data from different kinds of leukemia. We restricted the dataset to two conditions : chronic myeloid leukemia (74 samples) and healthy (76 samples). Furthermore, the dataset has also been reduced to 2500 genes. In the same way as for GSE6919, the set of undifferentiated modules is formed from random pathways while the set of differentiated modules is composed of pathways related to chronic myeloid leukemia.

3.3 Results and discussion

The precision-recall curves for both datasets are shown in Figure 2 and AUPR measures are reported in Table 2. We can observe from the GSE6919 curve that the degree scoring function performs best, followed by the mean-degree statistics. However, turning now our attention to the GSE13159 dataset, we can see that the baseline outperforms our approach. It is followed this time by the degree-variance scoring function. Hence, no single technique appears superior to others in all cases.

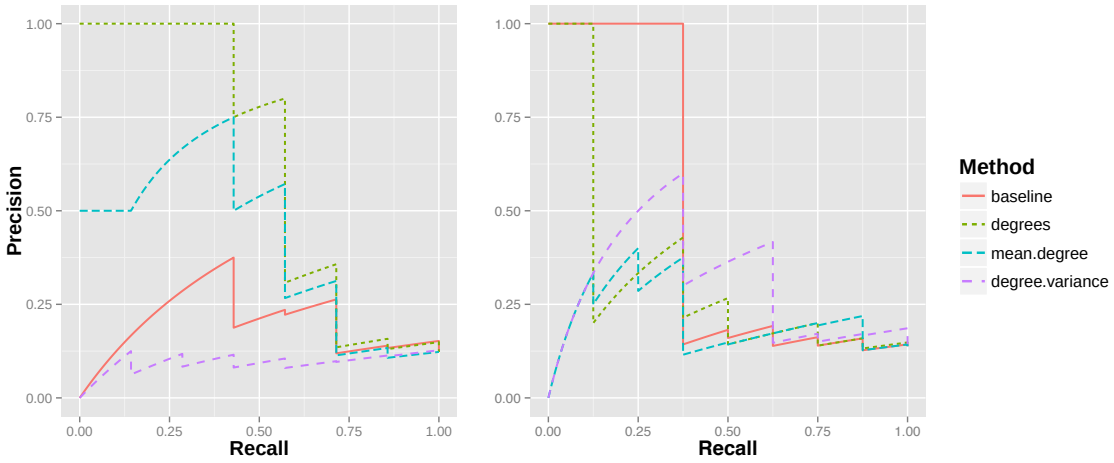
However, these results seem to underestimate the actual performances of our method. Indeed, if we consider for instance the most dysregulated pathways in prostate cancer according to the f_{degree} scoring function (as reported in Table 3), we can see that meaningful results are penalized by our initial definition of the differentiated modules. According to multiple studies [8,9], steroid hormones play a major role in human prostatic carcinogenesis. Besides, studies have shown associations between prostate cancer and alpha-linolenic acid [10]. Eventually, Brockhausen et al. [11] reports links between some kinds of *O*-glycans and adenocarcinomas from the prostate. Hence, pathways 4, 6 and 7 (in Table 3) are actually relevant to the disease, but are considered as false positives by the evaluation protocol.

Besides measuring AUPR performances, we also checked that the test behaves properly by testing the uniformity of the empirical distribution of p-values for undifferentiated modules. This has

Method	GSE6919	GSE13159
Baseline	0.20	0.40
Degrees	0.57	0.22
Mean degree	0.40	0.21
Degree variance	0.09	0.30

Table 2 – AUPR measures.

[1] Gene Expression Omnibus identifiers.



(a) GSE6919 **(b)** GSE13159

Fig. 2 – Precision-Recall curves

been done for the different scoring functions with a χ^2 test. None of these tests has shown enough evidence to reject the null hypothesis of uniformity. This result and a complementary visual inspection of the distribution indicate in particular a good control of type-I error.

4 Conclusions and perspectives

In this paper, we proposed a statistical framework to test the differential regulation of sets of genes. The primary contribution is the introduction of a computationally efficient permutation test. Indeed, this test does not require to reinfer GRNs (or recompute association measures for each pair of genes) for each permutation.

Promising results in terms of precision and recall has been obtained using very simple scoring functions. Besides testing the approach on additional datasets, there are plenty of opportunities for future works. Hitherto, we only considered scoring functions that rely on local properties of GRN topology. One might implement network statistics that takes long range dependencies into account. Furthermore, the data and procedure of evaluation are certainly a point to refine.

Rank	Pathway name
1	**MAPK signaling pathway**
2	**Prostate cancer**
3	**Apoptosis**
4	Steroid hormone biosynthesis
5	**Cytokine-cyt. receptor interaction**
6	Linoleic acid metabolism
7	Other types of O-glycan biosynthesis

Table 3 – Top-ranked pathways for GSE6919 dataset (prostate cancer) using f_{degree} scoring function. Pathways labeled as differentiated for the evaluation are in boldface.

References

1. Gill R. et al.: A statistical framework for differential network analysis from microarray data. BMC Bioinformatics 11(95) (2010)
2. Gambardella G. et al.: Differential network analysis for the identification of condition-specific pathway activity and regulation. Bioinformatics 29(14) (2013)
3. Liu, Y. et al.: Gene interaction enrichment and network analysis to identify dysregulated pathways and their interactions in complex diseases. BMC systems biology 6 (2012)
4. Amar, D. et al.: Dissection of regulatory networks that are altered in disease via differential co-expression. PLoS Comput. Biol. 9(3) (2013)
5. Meyer, P. E. et al.: minet: AR/Bioconductor package for inferring large transcriptional networks using mutual information. BMC bioinformatics 9(1) (2008)
6. Falcon, S. et al.: Hypergeometric testing used for gene set enrichment analysis. In Bioconductor case studies. Springer New York. (2008)
7. Coletta et al.: InSilico DB genomic datasets hub: an efficient starting point for analyzing genome-wide studies in GenePattern, Integrative Genomics Viewer, and R/Bioconductor. Genome Biology, 13(11) (2012)
8. Bosland, M. C.: The role of steroid hormones in prostate carcinogenesis. JNCI Monographs 2000(27) (2000)
9. Wilding, G.: The importance of steroid hormones in prostate cancer. Cancer surveys 14 (1991)
10. Azrad, M. et al.: Prostatic alpha-linolenic acid (ALA) is positively associated with aggressive prostate cancer: a relationship which may depend on genetic variation in ALA metabolism. PloS one, 7(12) (2012)
11. Brockhausen, I.: Pathways of *O*-glycan biosynthesis in cancer cells. Biochimica et Biophysica Acta (BBA)-General Subjects, 1473(1) (1999)

Outlier-Detecting Support Vector Regression for Modelling at the Transcriptome-Proteome Interface

Yawwani Gunawardana [1], Shuhei Fujiwara [3], Akiko Takeda [3], Christopher Woelk [2] and Mahesan Niranjan [1*]

[1]School of Electronics and Computer Science, University of Southampton, Southampton, UK
[2]Faculty of Medicine, Southampton General Hospital, University of Southampton, Southampton, UK
[3]Department of Mathematical Informatics,Graduate School of Information Science and Technology, University of Tokyo, Tokyo, Japan

Received on 2014; revised on XXXXX; accepted on XXXXX

Extented Abstract - MLSB

ABSTRACT

The mapping between high throughput measurements of gene expression, known as the transcriptome, and relative abundances of the corresponding proteins, known as the proteome, is not straingtforward due to different mechanisms of post-transcriptional and post-translational regulations taking place in cells. Yet, due to the relative ease with which transcriptome may be measured, mRNA levels are treated as proxies for protein levels. Our interest is in bridging the gap between these two levels of measurement. Specifically, we formulate a regression problem in which protein levels can be predicted from the corresponding mRNA levels and other proxies for efficiency of translation. Extending on previous work, where it was shown that outliers with respect to such regression are candidates for post-translational regulation, here we formulate a support vector regression problem in which a certain proportion of the data can be declared as outliers from the outset by defining a *clipped* loss function, *i.e.* very large errors are bounded by a threshold. Setting a threshold on the loss function is equivalent to defining a proportion of data as outliers, the latter being easier to detect outliers explicitly in the formulation of the problem. The resulting non-convex problem is solved by a difference of convex functions (DC) algorithm. On a data set of yeast transcriptome and proteome, assembled for this purpose and used previously, we show that the method is able to identify candidate post-translationally regulated proteins, confirmed by statistically significant enrichment of keywords of functional annotations.

Contact: mn@ecs.soton.ac.uk

1 INTRODUCTION

The mapping between high throughput measurements at the level of transcriptome and at the corresponding proteome is a complex one. While a large body of computational biology literature using advanced machine learning algorithms to transcriptomic data exist, it is acknowledged that the underlying biological function of interst happens more at the protein level and mRNA concentrations are seen as proxies for the corresponding protein concentrations. Several have measured mRNA and protein concentrations in the same biological samples and have attempted to show similarities between these two. Except under specific functional categories, correlation between the two is difficult to demonstrate. The reason for this is that different species of mRNA/proteins are regulated by different mechanisms at the post-transcriptional and post-translational levels.

The approach pursued in this work, starting from (Tuller *et al.*, 2007; Gunawardana and Niranjan, 2013), is to formulate a regression problem in which the response variable is the protein concentration and the covariates are the mRNA levels and other proxies for the stability and translation efficiency of the transcripts. Gunawardana and Niranjan (2013) used a linear regression model with a sparsity inducing regularizer (lasso) and showed that of about 37 features taken as inputs, a combination of mRNA levels and translation efficiencies (Greenbaum *et al.*, 2003) (experimentally measured polysome binding (Arava *et al.*, 2003) and sequence derived codon bias information (Wall *et al.*, 2005)) and can be reduced it to five dominant features to yield good prediction of protein levels. In fact, mRNA abundance, codon bian, tRNA adaptation index (tAI) , ribosome density and occupancy were selected as the most dominant five features. The outliers with respect to this linear regression were shown to carry significant over-representation of post-translationally regulated proteins, which is to be expected since the input covariates do not have any information about post-translational modifications.

In this contribution, we propose *an explicit formulation* for detecting post-translationally regulated proteins as outliers in a regression. This we believe is a much neater way of approaching the problem, than to simply implement a regression and hope the outliers to contain those post-translationally regulated genes. A particular aspect of this approach is that the proportion of data that should end up as outliers is a user tunable hyper-parameter. In practice, this may be derived from prior knowledge or be set from constraints such as the affordability of experimental work to confirm predictions of a computational model.

*to whom correspondence should be addressed

2 METHODS

In our previous work (Gunawardana and Niranjan, 2013), outliers were detected by looking at the regression plot of measured (P) and predicted protein concentrations ($\hat{P}$). In fact, the proteins lie furthest away from the regression line were considered as the outliers. However, in this paper we are going to use a novel approach to detect outliers. This technique is purely based on the loss function ($\ell(\boldsymbol{x}, y; \boldsymbol{w}, b)$) of the regression model. In this paper, we use the squared loss as $\ell(x, y; w, b)$ as the loss model. Proteins with largest losses (after ranking them according to the loss value) will be considered as the outliers. Clipped loss function is used to detect outliers in a robust manner (Yu *et al.*, 2010).

Suppose we have a set of m training samples $\{(x_i, y_i)\}_{i=1,\ldots,m}$ where $x_i \in \mathbb{R}^n$, $y_i \in \mathbb{R}$. Our goal is to predict y_i as $\hat{y} = \boldsymbol{w}^\top \boldsymbol{x} + b$ with small error. We define the *clipped loss function* as below:

$$\ell_U(\boldsymbol{x}, y; \boldsymbol{w}) := \min\{U, \ell(\boldsymbol{x}, y; \boldsymbol{w}, b)\}$$

using a hyper parameter $U > 0$ to denote the clipping position.

2.1 Truncated Loss Model

Following is the model for our regression problem:

$$\min_{\boldsymbol{w},b} \sum_i \ell_U(\boldsymbol{x}_i, y_i; \boldsymbol{w}, b) + \lambda \|\boldsymbol{w}\|^2,$$

where $U > 0$ and $\lambda > 0$ are hyper parameters. It is troublesome to control U to define outliers. Therefore, we use a parameter which corresponds to the outlier ratio $\mu \in [0, 1)$ instead of U and consider the outlier-detecting regression model:

$$\begin{aligned} \min_{\boldsymbol{w},b,\boldsymbol{\eta}} \quad & \frac{1}{(1-\mu)m} \sum_i \eta_i \ell(\boldsymbol{x}_i, y_i; \boldsymbol{w}, b) + \lambda \|\boldsymbol{w}\|^2 \\ \text{s.t.} \quad & \textstyle\sum_i (1-\eta_i) \le \mu m, \ \ 0 \le \eta_i \le 1, \ \forall i, \end{aligned} \tag{1}$$

where $\mu \in [0, 1)$ and $\lambda \in (0, \infty)$ are hyper parameters.

Note that $\sum_i (1 - \eta_i) = \mu m$ holds at the optimality. The samples (x_i, y_i) with $\eta_i^* = 0$ can be regarded as an outlier for small $\mu > 0$. However, this is a non-convex problem and finding a gobal solution for a non-convex problem is very difficult.

2.2 Difference of Convex Functions (DC) Algorithm

Difference of Convex Functions (DC) algorithm is a mathematical technique to find plausible solutions for non-convex functions (Pham Dinh and Le Thi, 1997). Thus, we use this method to solve the non-convex clipped function of the regression model (1). Steps of minimizing the difference of convex functions are explained in the following section.

We describe the objective function of regresion model (1) by using a difference of convex functions and rewrite it as

$$\begin{aligned} \min_{\boldsymbol{w},b} \quad & \frac{1}{(1-\mu)m} \left\{ \sum_i \ell(\boldsymbol{x}_i, y_i; \boldsymbol{w}, b) - \mu m \phi_{1-\mu}(\boldsymbol{w}, b) \right\} + \lambda \|\boldsymbol{w}\|^2, \\ = \quad & \underbrace{\frac{1}{(1-\mu)m} \sum_i \ell(\boldsymbol{x}_i, y_i; \boldsymbol{w}, b) + \lambda \|\boldsymbol{w}\|^2}_{\text{convex}} - \underbrace{\frac{\mu}{1-\mu} \phi_{1-\mu}(\boldsymbol{w}, b)}_{\text{convex}}, \end{aligned} \tag{2}$$

where $\phi_{1-\mu}(\boldsymbol{w}, b)$ is $(1-\mu)$-Conditional Value-at-Risk (CVaR), known as a popular financial risk measure, which can be described as

$$\phi_{1-\mu}(\boldsymbol{w}, b) := \min_\alpha \ \alpha + \frac{1}{\mu m} \sum_{i=1}^m [\ell(\boldsymbol{x}_i, y_i; \boldsymbol{w}, b) - \alpha]^+, \tag{3}$$

using Theorem 10 in Rockafellar and Uryasev (2002), or equivalently as

$$\begin{aligned} \phi_{1-\mu}(\boldsymbol{w}, b) = \quad & \max_{\boldsymbol{\eta}} \frac{1}{\mu m} \sum_i (1-\eta_i) \ell(\boldsymbol{x}_i, y_i; \boldsymbol{w}, b) \\ & \text{s.t.} \textstyle\sum_i (1-\eta_i) = \mu m, \ \ 0 \le \eta_i \le 1, \ \forall i. \end{aligned}$$

We define the set of outliers by Θ using the optimal solution η^* of (1):

$$\Theta := \{i \in \{1, \ldots, m\} : \eta_i^* < 1\}$$

The loss function in (2) is written as

$$\frac{1}{(1-\mu)m} \left\{ \sum_{i=1}^m \ell(\boldsymbol{x}_i, y_i; \boldsymbol{w}, b) - \sum_{i \in \Theta} \ell(\boldsymbol{x}_i, y_i; \boldsymbol{w}, b) \right\}.$$

Difference of convex functions algorithm sequentially linearizes the concave part of (2) and solves the convex subproblem. Let (w_k, b_k) be the solution obtained in the $(k-1)$th iteration. In the kth iteration, we solve the following subproblem:

$$\min_{\boldsymbol{w},b} \lambda \|\boldsymbol{w}\|^2 + \frac{1}{(1-\mu)m} \left\{ \sum_i \ell(\boldsymbol{x}_i, y_i; \boldsymbol{w}, b) - \mu m (\boldsymbol{g}_w^{k\top} \boldsymbol{w} + g_b^k b) \right\} \tag{4}$$

where $\boldsymbol{g}_w^k \in \partial_w \phi_{1-\mu}(\boldsymbol{w}_k, b_k)$ and $g_b^k \in \partial_b \phi_{1-\mu}(\boldsymbol{w}_k, b_k)$ are a subgradient of $\phi_{1-\mu}(\boldsymbol{w}, b)$ at $(\boldsymbol{w}_k, b_k)$ which can be calculated by sorting the loss $\ell_i(x_i, y_i; w_k, b_k)$.

The sequence $\{(\boldsymbol{w}_k, b_k)\}$ generated by equation (4) has the following good convergence properties: The objective value is decreasing and every limit point of the sequence is a critical point defined in Pham Dinh and Le Thi (1997) of (2). The critical point is also called *generalized KKT point* which is a necessary condition of a local solution.

Algorithm 1 shows the pseudo code of the DC regression model algorithm which was developed using the `CVX` package in `MATLAB` environment.

Algorithm 1 DC Algorithm for Linear Regression

Require: (λ, μ), $\boldsymbol{\alpha} = (\boldsymbol{w}_0, b_0)$, $\boldsymbol{X}$ input matrix, $\boldsymbol{Genes}$ is an array of gene names of the samples and $\boldsymbol{y}$ output vector
$\boldsymbol{\eta}(i)_{i=1}^{N} = 1$ % binary vector to represent outliers
$\boldsymbol{D}$ = diag($\boldsymbol{\eta}$)
$\boldsymbol{D_h}$ = sqrt($\boldsymbol{D}$)
$k = 0$
repeat

// given $\boldsymbol{D}$ compute $\boldsymbol{\alpha}$
$\min_{\boldsymbol{\alpha}}\{\|\boldsymbol{D_h X\alpha} - \boldsymbol{D_h y}\|^2 + \lambda\|\boldsymbol{\alpha}\|^2\}$

// given $\boldsymbol{\alpha}$ compute $\boldsymbol{D}$
$\boldsymbol{E}$ = $\{\boldsymbol{X\alpha} - \boldsymbol{y}\}^2$ // compute squared loss
ind = sort($\boldsymbol{E}$) // sort the samples based on error
$outierInd$ = ceil(μm:m) // get the indicies of the largest μ errors
$\boldsymbol{\eta}(outierInd)$ = 0; // make outlier samples to 0
$\boldsymbol{D}$ = diag($\boldsymbol{\eta}$)
$\boldsymbol{D_h}$ = sqrt($\boldsymbol{D}$)

until $k < 100$
$outliers$ = $Genes(outierInd)$ // obtain the final outlier set

2.3 Post-translational Regulation Annotation Check

Similar to the Gunawardana and Niranjan, 2013's study, functional annotation check was carried out in two levels (*i.e.* coarse level and finer level). All the annotation databases are similar to the previous study.

3 RESULTS

Five main features at the transcriptome level (*i.e.* mRNA abundance, ribosome occupancy (Greenbaum *et al.*, 2003), ribosome density (Arava *et al.*, 2003), tAI and codon bias (Wall *et al.*, 2005)) of *Saccharomyces cerevisiae* organism were selected as the inputs for the regression model (using arsity inducing lasso) and the respective protein abundances (Wang *et al.*, 2012) were used as the output.

In previous work (Gunawardana and Niranjan, 2013), outliers were detected by looking at the regression plot of measured (P) and predicted protein concentrations ($\hat{P}$). In fact, proteins found furthest away from the regression line were considered as the outliers. We selected 50 proteins which found beyond the 97.5% prediction boundary as the baseline set of outliers. Coarse level annotation check (looking for only PTMs key words) gave p-value < 0.02 and finer level functional annotations which contained PTMs coupled with motif information (*i.e.* Phosphorylation + PEST motifs, Acetylation + N-termini segments and Ubiquitination + D or KEN Box motifs) gave p-value $< 2.1078 \times 10^{-10}$. Both these p-values are statistically significant and provide high confidence levels for the over-representation of post-translational regulations (PTR) in outlier set.

3.1 DC Algorithm PTM Outliers

Same functional annotation checks was carried out with the new 50 outliers detected by the DCA (setting μ to 97.5%). Forty proteins were found with PTM key word at the coarse level providing a p-value < 0.048. Thirty three proteins were detected at the finer level annotation check (PTMs+motif) with a high confidence level of p-value $< 4.4977 \times 10^{-06}$. Though these confidence levels are lower than our previous method outliers, these p-values can also be considered as high confidence levels with respect to the confidence level threshold ($p < 0.05$) of accepting a hypothesis in biomedical research (McDonald, 2009). Thus, the outliers detected by DC algorithm are also highly enriched with post-translational regulations. Figure 1 shows these outliers in a scatter plot of predicted versus true protein concentrations.

3.2 Analysis of Two Outlier Sets

We compared the two outliers sets (previous method and DC algorithm) and found 20 genes common to both sets. Fifteen proteins out of 20 had PTMs with motif information and the corresponding confidence level is p-value $< 2.6104 \times 10^{-05}$. We subjected the rest of the proteins (those are not common between two sets) in to GO enrichment analysis using Gene Ontology Enrichment Analysis Software Toolkit (GOEAST) (Zheng and Wang, 2008) . We observed that non common genes in new outlier set enriched with biosynthesis processes. Ten genes were identified among four biosynthesis processes (*i.e.* cellular amino acid biosynthesis, organic acid biosynthesis, carboxylic acid biosynthesis and organonitrogen compound biosynthesis). Non common genes in the old outlier set (Gunawardana and Niranjan, 2013) were enriched with 15 ribosomal properties and five biosynthesis process (different from new outlier set biosynthesis processes). Twenty two genes were identifies as ribosomal proteins from the non-common genes in the privous work (Gunawardana and Niranjan, 2013) outlier set.

We also noted that 15 outliers detected from the DC algorithm were found in the lower region of the regression plot (measured abundance $P >$ predicted abundance $\hat{P}$). However, in our main study we considered that the protein degradation by post-translation modifications reduces the measured protein abundance with respect to the actual abundances. Therefore, we looked in the upper region ($P < \hat{P}$) to detect outliers. However, in this method that 15 outliers were found in the lower region ($P < \hat{P}$). We note that seven proteins did not show any post-translational regulations (PTM + motif information). Thus, the high number of lower region protein selection by DC algorithm caused the total reduction of PTR detection with respect to the previous method.

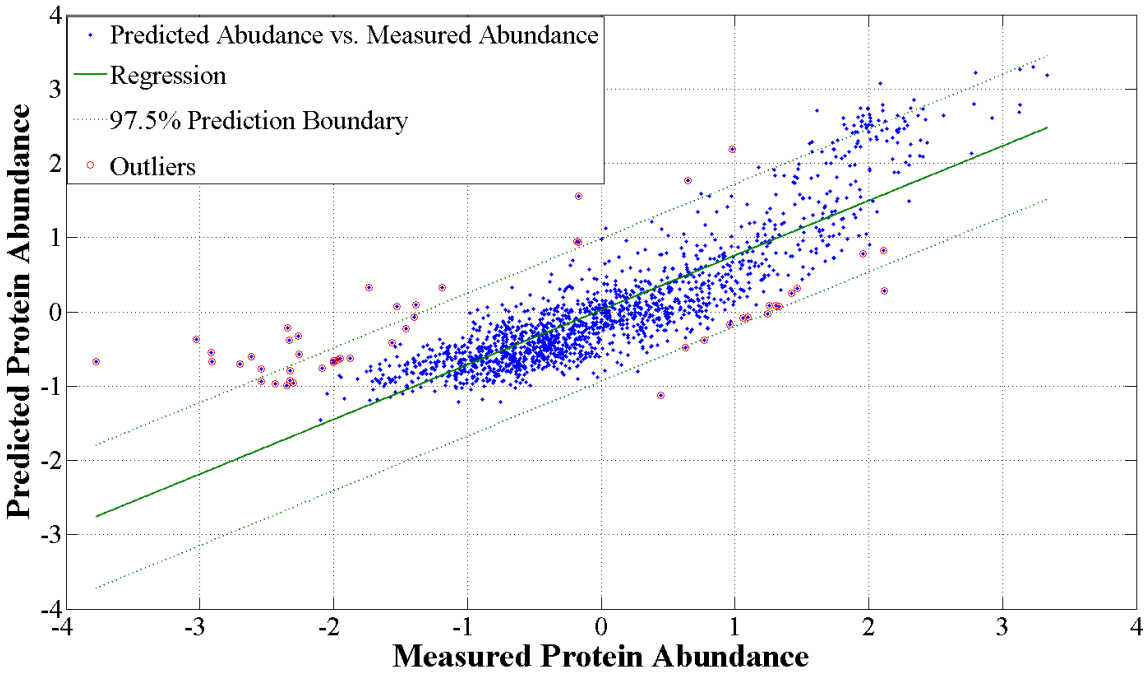


Fig. 1. Outlier detection by DC Algorithm : Least accurate 50 outliers are shown in red circles

4 CONCLUSION

In this work, we have developed a computational approach to detect post-translationally regulated proteins as outliers in a regression of protein levels as functions of mRNA levels and efficiencies of translation. The significant novelty in the work is the *explicit* formulation of a fraction of the data to be treated as outliers in setting up a support vector regression problem. This non-convex optimization problem, which we solve by a DC programming approach, is shown to identify post-translationally regulated yeast proteins, confirmed by functional annotations of gene ontology keywords.

REFERENCES

Arava, Y., Wang, Y., Storey, J. D., Liu, C. L., Brown, P. O., and Herschlag, D. (2003). Genome-wide analysis of mRNA translation profiles in Saccharomyces cerevisiae. *Proceedings of the National Academy of Sciences*, **100**(7), 3889–3894.

Greenbaum, D., Colangelo, C., Williams, K., and Gerstein, M. (2003). Comparing protein abundance and mRNA expression levels on a genomic scale. *Genome Biology*, **4**(9), 117.

Gunawardana, Y. and Niranjan, M. (2013). Bridging the gap between transcriptome and proteome measurements identifies post-translationally regulated genes. *Bioinformatics*, **29**(23), 3060–3066.

McDonald, J. H. (2009). *Basic concepts of hypothesis testing*, volume 2.

Pham Dinh, T. and Le Thi, H. A. (1997). Convex analysis approach to d.c. programming: Theory, algorithms and applications. *Acta Mathematica Vietnamica*, **22**(1), 289–355.

Rockafellar, R. T. and Uryasev, S. (2002). Conditional value-at-risk for general loss distributions. *Journal of Banking & Finance*, **26**(7), 1443–1472.

Tuller, T., Kupiec, M., and Ruppin, E. (2007). Determinants of protein abundance and translation efficiency in S.Cerevisiae. *PLoS Computational Biology*, **3**(12), e248.

Wall, D. P., Hirsh, A. E., Fraser, H. B., Kumm, J., Giaever, G., Eisen, M. B., and Feldman, M. W. (2005). Functional genomic analysis of the rates of protein evolution. *Proceedings of the National Academy of Sciences of the United States of America*, **102**(15), 5483–5488.

Wang, M., Weiss, M., Simonovic, M., Haertinger, G., Schimpf, P., Hengartner, M., and Mering, C. (2012). Paxdb, a database of protein abundance averages across all three domains of life. *Molecular and cellular proteomics : MCP*, **11**(8), 492–500.

Yu, Y., Yang, M., Xu, L., White, M., and Schuurmans, D. (2010). Relaxed clipping: A global training method for robust regression and classification. In *Neural Information Processing Systems*, pages 2532–2540.

Zheng, Q. and Wang, X.-J. (2008). Goeast: a web-based software toolkit for gene ontology enrichment analysis. *Nucleic acids research*, **36**(suppl 2), W358–W363.

A hybrid approach for the inference and modelling of gene regulatory networks

Vân Anh Huynh-Thu[1] and Guido Sanguinetti[1,2]

[1] School of Informatics, University of Edinburgh, UK
[2] SynthSys - Systems and Synthetic Biology, University of Edinburgh, UK

1 Introduction

An important open problem in computational systems biology is the reconstruction of topologies of gene regulatory networks (GRNs) using high throughput genomic data, in particular gene expression data. Among existing GRN inference algorithms, one can find on one side methods that compute statistical dependencies, such as mutual information or partial correlation, between the expression patterns of all pairs of genes. These methods usually scale with the number of genes and can therefore infer large GRNs very efficiently. However, since they are essentially model-free, the networks learned with these methods can not be used to make predictions of gene expression profiles under novel experimental conditions. On the other side, model-based methods attempt to capture the dynamics of the system under study. These models are typically based on systems of ordinary or stochastic differential equations and can generate realistic behaviour. They however usually include many unknown parameters, and are therefore limited to small networks. Here, we propose a new hybrid approach that combines formal dynamical modelling with the efficiency of a model-free method, allowing to reconstruct the topologies of networks of hundreds of genes. We present results on artificial time series expression data, showing that our method is able to capture observed and latent dynamics, and is competitive with existing GRN inference approaches.

2 Methodology

2.1 Gene expression model

At the heart of our framework, we use the *on/off model* of gene expression [8], a simple, yet plausible, model where the rate of transcription of a gene can vary between two levels depending on the activity state μ of the promoter of the gene. The expression x of a gene is modelled through the following stochastic differential equation:

$$\mathrm{d}x_i = (A_i\mu_i(t) + b_i - \lambda_i x_i)\mathrm{d}t + \sigma\mathrm{d}w(t), \tag{1}$$

where subscript i refers to the i^{th} target gene. Here, μ_i is a binary variable (promoter is either active or inactive), whose state depends on the expression of the transcription factors that bind to the promoter. $\Theta = \{A_i, b_i, \lambda_i\}$ is the set of kinetic parameters. A_i represents the efficiency of the promoter in recruiting polymerase when being in the active state. The sign of A_i defines the type

of regulation: either activation or repression. b_i represents the basal transcription rate and λ_i is the exponential decay constant of x_i. The term $\sigma \mathrm{d}w(t)$ represents a white noise-driving process with variance σ^2.

In our framework, we model gene expression $x_i(t)$ as a Gaussian process [9]. One important advantage of using Gaussian processes is that various probability distributions can be computed exactly. The Gaussian process $x_i(t)$ is completely defined by its mean $m_i(t)$ (which depends on the promoter state $\mu_i(t)$) and covariance $k_i(t, t')$, and we assume that we observe this process with i.i.d. Gaussian noise: $\hat{x}_i \sim \mathcal{N}(x_i, \sigma^2_{obs})$, where σ^2_{obs} is the variance of the observation noise. Given the trajectory of the promoter state $\mu_i(t)$, one can compute $m_i(t)$ using Equation (1), as well as the marginal log-likelihood of the observations, given by:

$$\log \mathcal{L} = -\frac{N}{2}\log(2\pi) - \frac{1}{2}|K_i + \sigma^2_{obs} I| - \frac{1}{2}(\hat{\mathbf{x}}_i - \mathbf{m}_i)^\top (K_i + \sigma^2_{obs} I)^{-1}(\hat{\mathbf{x}}_i - \mathbf{m}_i), \tag{2}$$

where N is the number of observation time points, $\mathbf{m}_i$ is a vector containing the values of $m_i(t)$ at these time points, K_i is the covariance matrix, and I is the identity matrix.

Within this context, our goal is, for each target gene, (a) to identify the trajectory $\mu_i(t)$ that maximizes the marginal log-likelihood, and (b) to identify the regulators of the target gene, i.e. the transcription factors that influence $\mu_i(t)$.

2.2 Network reconstruction

Tree-based methods have been applied successfully in the inference of GRNs [5]. These methods have appealing properties: they are non-parametric, can deal with high-dimensional datasets, and are highly scalable. Therefore, we chose to resort to decision trees in order to infer the promoter state $\mu_i(t)$ of each target gene, from the expression levels of the candidate regulators. However, since in our case $\mu_i(t)$ is a latent variable, we propose a new variant of the decision tree algorithm. In this variant, the tree is constructed top-down, starting from a root node $\mathcal{N}$ where we assume that $\mu_i^{\mathcal{N}}(t) = 0, \forall t$, with corresponding log-likelihood $\log \mathcal{L}(\mu_i^{\mathcal{N}})$. Given the observed expression $\hat{\mathbf{y}}$ of a candidate regulator and a cut-point c, a candidate trajectory $\mu_i^{y,c}(t)$ is obtained using:

$$\mu_i^{y,c}(t_k) = \begin{cases} 0, & \text{if } \hat{y}(t_k) < c, \\ 1, & \text{if } \hat{y}(t_k) \geq c, \end{cases} \tag{3}$$

for each observation time point t_k. Between two time points, values of $\mu_i^{y,c}(t)$ are merely set to the value obtained at the previous time point. The best candidate regulator and the best cut-point are then chosen, i.e. those that most increase the log-likelihood:

$$d(\mathcal{N}) = \log \mathcal{L}(\mu_i^{y,c}) - \log \mathcal{L}(\mu_i^{\mathcal{N}}). \tag{4}$$

Let $\mu_i^{y,c*}(t)$ be the trajectory that maximizes $d(\mathcal{N})$. Two child nodes $\mathcal{N}_0$ and $\mathcal{N}_1$ are created and the same procedure is applied to refine $\mu_i^{y,c*}(t)$. In node $\mathcal{N}_0$ (resp. $\mathcal{N}_1$), values of $\mu_i^{y,c*}(t)$ are refined

at the time points where $\mu_i^{y,c*}$ is equal to 0 (resp. 1). A node $\mathcal{N}_0$ (resp. $\mathcal{N}_1$) becomes a terminal node if the log-likelihood can not be increased, and this node contains 0 (resp. 1) as predicted value for μ.

To avoid an over-fitting of the observed data, an ensemble of randomized trees can be constructed, e.g. by randomizing the selection of the cut-point as in the Extra-Trees algorithm [4]. The prediction of $\mu_i(t)$ is then averaged over the different trees, yielding a probability for the promoter state to be active at time t.

An interesting property of decision trees is that it is possible to derive from the learned tree-based model an importance measure that allows to rank the candidate regulators according to their relevance for predicting $\mu_i(t)$ and thereby the expression of the target gene. For a single tree, the importance value of a candidate regulator is taken as the sum of the $d(\mathcal{N})$ values (4) obtained at all the tree nodes where this regulator was selected. For an ensemble of trees, the importances are averaged over all individual trees. The importance of a candidate regulator in the prediction of the target gene expression pattern is taken as an indication of a putative regulatory link.

3 Results

3.1 Toy data

As a first validation, we used 5 artificial networks of 100 genes, and generated synthetic expression data using a switch model based on Equation (1). For each network, the observation data consist of 10 independent time series of 21 time points. Modelling results are shown in Figure 1 for one gene. We notice that our approach provides a good prediction of the promoter state, as well as a good fitting of the gene expression.

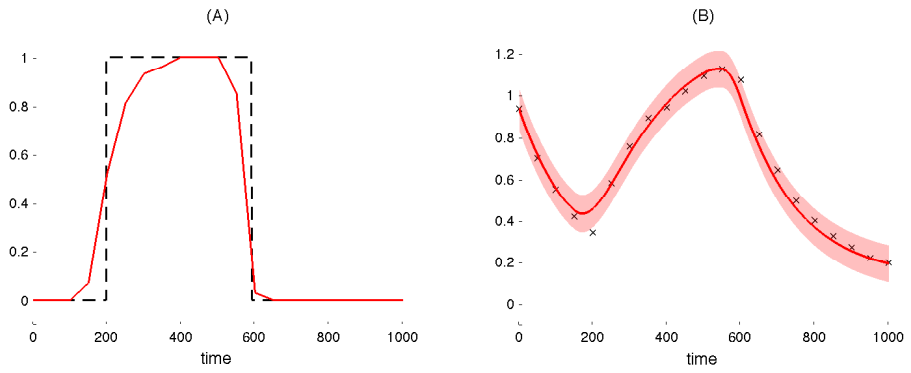


Fig. 1. Modelling results on toy data, for one target gene. (A) Predicted promoter state $\mu(t)$ (solid red) versus true state (dashed black). (B) Posterior mean of gene expression $x(t)$ (solid red), with confidence intervals. Data points $\hat{\mathbf{x}}$ are shown as black crosses.

Next, we checked if our method was able to correctly learn the network topologies, and we compared it to other existing GRN inference procedures: time-lagged variants of GENIE3 [5] and CLR [3], simone [2], and G1DBN [6]. Our method ("Jump trees") yields the highest area under the precision-recall curve (AUPR), as shown in Figure 2. Table 1 ("Toy" column) indicates the AUPR values, averaged over the 5 networks.

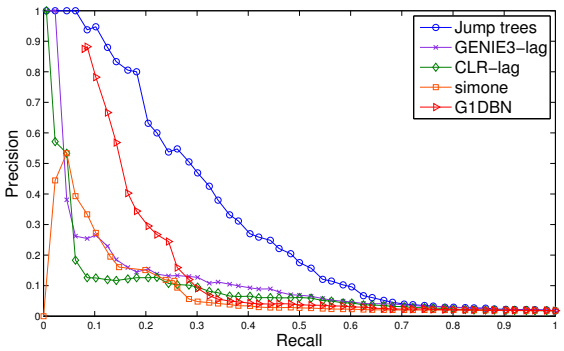


Fig. 2. Precision-recall curves for one 100-gene network.

	Toy	DREAM4
Jump trees	**0.261 ± 0.057**	**0.183 ± 0.061**
GENIE3-lag	0.119 ± 0.009	0.180 ± 0.056
CLR-lag	0.092 ± 0.009	0.172 ± 0.047
simone	0.076 ± 0.022	0.072 ± 0.022
G1DBN	0.143 ± 0.038	0.121 ± 0.046

Table 1. Comparison of network inference methods (mean AUPR and standard deviation).

3.2 DREAM4 challenge

The Dialogue for Reverse Engineering Assessments and Methods (DREAM) project provides benchmarks for the evaluation of GRN inference algorithms [7]. We applied the different methods to infer the 100-gene networks of the DREAM4 challenge, exploiting time series data only. Results are shown in Table 1 ("DREAM4" column). As expected, the performance of our method decreases compared to the one obtained on the toy data, due to the mismatch between our model and the one used to simulate the DREAM4 data. However, the AUPR remains higher than those of the other procedures.

4 Conclusions

We propose a new hybrid approach for the inference and dynamical modelling of gene regulatory networks. Preliminary results show that this approach performs well on artificial data. Further steps include an extension of the method that would allow a control on the number of switches in the promoter state trajectory, the evaluation of the method when predicting expression profiles under new conditions, as well as its application for the inference of real networks.

References

1. L. Breiman, J. H. Friedman, R. A. Olsen, and C. J. Stone. *Classification and Regression Trees*. Wadsworth International (California), 1984.
2. C. Charbonnier, J. Chiquet, and C. Ambroise. Weighted-lasso for structured network inference from time course data. *Statistical Applications in Genetics and Molecular Biology*, 9(1):Article 15, 2010.
3. J. J. Faith, B. Hayete, J. T. Thaden, I. Mogno, J. Wierzbowski, G. Cottarel, S. Kasif, J. J. Collins, and T. S. Gardner. Large-scale mapping and validation of *Escherichia coli* transcriptional regulation from a compendium of expression profiles. *PLoS Biology*, 5(1):e8, 2007.
4. P. Geurts, D. Ernst, and L. Wehenkel. Extremely randomized trees. *Machine Learning*, 36(1):3–42, 2006.
5. V. A. Huynh-Thu, A. Irrthum, L. Wehenkel, and P. Geurts. Inferring regulatory networks from expression data using tree-based methods. *PLoS ONE*, 5(9):e12776, 2010.
6. S. Lèbre. Inferring dynamic bayesian networks with low order independencies. *Statistical Applications in Genetics and Molecular Biology*, 8(1):Article 9, 2009.
7. D. Marbach, J. C. Costello, R. Küffner, N. Vega, R. J. Prill, D. M. Camacho, K. R. Allison, the DREAM5 Consortium, M. Kellis, J. J. Collins, and G. Stolovitzky. Wisdom of crowds for robust gene network inference. *Nature Methods*, 9:796–804, 2012.
8. M. Ptashne and A. Gann. *Genes and Signals*. Cold Harbor Spring Laboratory Press, New York, 2002.
9. C. Rasmussen and C. K. I. Williams. *Gaussian Processes for Machine Learning*. MIT Press, 2006.

Binary classification of metagenomic samples using discriminative DNA superstrings

Karel Jalovec[1] and Filip Železný[1]

[1]CTU in Prague FEE Dept. of Computer Science and Engineering

Motivation

Increasing amount of data obtained by the NGS technologies increases the urge of effective analysis of this data. This work presents a tool for binary classification of metagenomic samples. Metagenomic samples consist of a large amount of short DNA strings (also called reads), which belong to different organisms present in an environment from which the sample was taken. Behavior of an environment can be affected by the contamination by the organisms, which originaly do not belong in this environment. The goal of this work is to develop a classification method based on DNA superstrings that can accurately classify metagenomic samples. Classifiers obtained by this method can be used for determining whether newly obtained metagenomic samples are contaminated (positive) or clean (negative) without the need of identification of particular organisms present in the sample. We want to achieve this goal by establishing a modified sequence assembly task for finding the most discriminatory DNA superstrings.
We assume that standard a approach for this kind of analysis would be to assemble all the samples and try to find the most discriminatory motifs. Both tasks are very computationally demanding. Our method should solve both these tasks simultaneously.

Related problems and preliminaries

Shortest common superstring - SCS
Given a set $S = \{s_1, s_2, ..., s_n\}$ of short strings s_i, determine the shortest string R such that each string $s_i \in S$ is a substring of R.

Consistent superstring - CSS
The most common variants are LCSS (Longest CSS) and SCSS (Shortest CSS). Given a set of positive strings $P = \{p_1, p_2, ..., p_n\}$ and a set of negative strings $N = \{n_1, n_2, ..., n_m\}$, determine a (longest/shortest) superstring R such that each positive string $p_i \in P$ is a substring of R and each negative string $n_j \in N$ is not a substring of R.

Suffix-prefix concatenation
This function (denoted by the $\circ$ symbol) takes two strings as an input. If a suffix of one string matches a prefix of the second string, this operation concatenates the unmatched portion of the second string to the end of the first string. In case of multiple possible concatenations, it is usually considered only the solution with the longest overlap.

String coverage
String r covers certain percentage of string s when r is a substring of s. The portion of symbols of string s matched by symbols of string r are covered. This function can be extended to measure the percentage of string s which is covered by a set of strings $R = \{r_1, r_2, \ldots, r_n\}$.

Problem definition

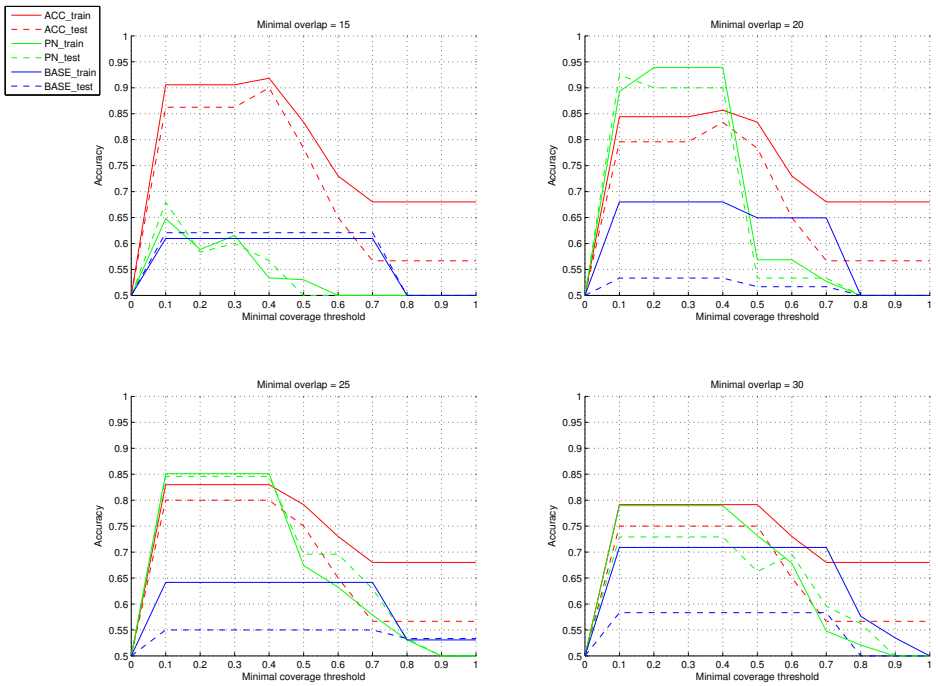
Consider two multisets of reads $S^N = \{\{s_1, s_2, \ldots, s_n\}\}$, $S^P = \{\{s_1, s_2, \ldots, s_m\}\}$ and two powersets 2^{S^N}, $2^{(S^P \cup S^N)}$. Metagenomic sample S_a can be either positive (contaminated) $S_a^+ \in 2^{(S^P \cup S^N)}$ or negative (clean) $S_a^- \in 2^{S^N}$. We are presented by a training set of positive samples $S^+ = \{S_1^+, S_2^+, \ldots, S_k^+\}$ and a set of negative samples $S^- = \{S_1^-, S_2^-, \ldots, S_l^-\}$. The goal is to find a permutation p such that $p = p_1, p_2, \ldots, p_{|p|}$ and $p_{i \in 1..|p|} \in S^P$. Superstring λ_p can be created from permutation p by using the operation suffix-prefix concatenation $\lambda_p = p_1 \circ p_2 \circ \ldots \circ p_{|p|}$. A trained classifier consists of a superstring and a coverage threshold value Θ. The classifier computes coverage of its superstring from the reads contained within the sample and compares it to its coverage threshold. If the computed coverage is higher than the coverage threshold value, the sample is marked as positive. If the coverage is insufficient, the sample is marked as negative.

Methods

To obtain the optimal solution, it would be necessary to search all possible permutations of reads from all the samples. This would be nearly impossible given the fact, that a single metagenomic sample can contain tens of millions of reads. To prune the search space we use a graph-based approach that is very commonly used for sequence assembly tasks. We construct an overlap graph formed by all strings from all samples. To speed up the construction process, we use the FM-index structure[1]. Each vertex in the graph is annotated by the PN-measure value. PN-measure value of the read is computed as a difference between its positive and negative occurrence counts. More positive occurrences of the read result in a higher PN-measure value. We propose to use beam search to discover the most discriminatory paths in the graph.
We have developed three types of classifiers. The first classifier (BASE) serves as a baseline solution. All samples were assembled using the ReadJoiner[2] assembler in the first stage and in the second stage, the most discriminatory motif was found using the MEME toolkit[3]. Both the second (PN) and the third (ACC) classifiers use beam search to find the most discriminatory paths in an overlap graph during the training stage. Initial beam of the algorithm is filled with elementary paths that consist of a single vertex. Vertices for elementary paths are selected with respect to their PN-measure values. The algorithm selects the most discriminatory path in the beam in each step and tries to extend it. Path extension adds new vertices to the left, to the right or to the both sides of the path. If there is more than one extension available all of them are added into the beam. Original path is marked as inextensible and also kept in the beam. The reason for keeping the original path in the beam is that the original path can be more discriminatory than all its extensions. Marking the path as inextensible prevents the algorithm from recurring extension of the previously extended path. The algorithm stops when all the paths in the beam are marked as inextensible. The PN and the ACC classifiers differ in the path feasibility evaluation process. The PN classifier computes cumulative PN-measure value of the path. The PN-measure values of the reads along the path are summed up and the higher the cumulative PN-measure value is, the preferable the path is. This leads to a faster but less accurate search. The ACC classifier evaluates the feasibility of each path in the same way as the classification procedure itself works. Extended path is converted into the superstring by concatenation of its reads and the accuracy value of this superstring is computed. This leads to a slower but more accurate search.

Results

We tested the performance of our classifiers with a simulated data set. We randomly selected four organisms for our simulated environment and created a certain amount of samples by individually sequencing all the genomes and putting them together. Positive samples were contaminated with an additional organism. We run our algorithm with four different values of a minimal overlap threshold for creating an overlap graph. The performance of the classification is depicted in the image below. The beam width has fixed value of 15.



Accuracy of classifiers w.r.t minimal coverage threshold

The time requirements of the algorithm with respect to the amount of data are described below. Learning was performed on two training sets consisting of 90k and 150k reads. Learning from the smaller dataset took the algorithm between 5 and 10 seconds for the ACC classifier and about 2 seconds for the PN classifier. Larger dataset required 15 to 80 seconds for the ACC classifier and about 5 seconds for the PN classifier. Learning of the baseline classifier using the meme toolkit requires a time constant for which the tool searches for discriminatory motifs. We selected a constant with a value of 300 seconds. Classification procedure requires miliseconds for classifying a single sample consisting of 30k to 40k reads.

Future improvements

So far, our method is able to find only a single discriminatory superstring. Our plan is to extend the method so that each classifier can contain more than a single discriminatory superstring which will lead to a better classification accuracy. Next improvement we plan to incorporate in our method is to increase the biological interpretability of the discriminatory superstrings. So far we did not pay attention whether the superstring is a biologicaly valid genomic sequence.

References

[1] Simpson and Durbin: **Efficient construction of an assembly string graph using the FM-index**. Bioinformatics 26:i367–i373, 2010, http://dx.doi.org/10.1093/bioinformatics/btq217.

[2] Gonnella and Kurtz: **Readjoiner: a fast and memory efficient string graph-based sequence assembler**. BMC Bioinformatics 2012 13:82.

[3] T. L. Bailey et al.: **MEME SUITE: tools for motif discovery and searching**. Nucleic Acids Research, 2009, Vol. 37, Web Server issue doi:10.1093/nar/gkp335

Acknowledgments

- This work has been supported by the Czech Science Foundation project P202/12/2032.

- Access to computing and storage facilities owned by parties and projects contributing to the National Grid Infrastructure MetaCentrum, provided under the programme "Projects of Large Infrastructure for Research, Development, and Innovations" (LM2010005), is greatly appreciated.

BIOINFORMATICS

Vol. 00 no. 00 2014
Pages 1–4

Parameter Estimation in Computational Biology by Approximate Bayesian Computation coupled with Sensitivity Analysis

Xin Liu and Mahesan Niranjan

School of Electronics and Computer Science, University of Southampton, Southampton, UK

Received on XXXXX; revised on XXXXX; accepted on XXXXX

Extented Abstract - MLSB

ABSTRACT

We address the problem of parameter estimation in models of systems biology from noisy observations of model outputs. The models we consider are characterized by simultaneous deterministic nonlinear differential equations whose parameters are either taken from *in vitro* experiments, or are hand-tuned during the model development process to reproduces observations. We consider the family of algorithms coming under the Bayesian formulation of Approximate Bayesian Computation (ABC) and show that sensitivity analysis could be deployed to quantify the relative roles of different parameters in the system. Parameters to which a system is relatively less sensitive (known as sloppy parameters) need not be estimated to high precision, while the values of parameters that are more critical (stiff parameters) need to be determined with care.

The difficulty in estimating problem in high dimensions suggests a systematic re-allocation of computational effort, and we propose a three stage strategy in which sloppy parameters of a model are estimated in a coarse search followed by re-estimation of the stiff parameters to tighter error tolerances. We demonstrate the effectiveness of the proposed method on three oscillatory models and one transient response model taken from the systems biology literature.

Contact: mn@ecs.soton.ac.uk

1 INTRODUCTION

Computational modeling of biological systems is about describing quantitative relationships of biochemical reactions by systems of differential equations (Kitano, 2002). Knowledge of biological processes captured in such equations, when solutions to them match measurements made from the system of interest, help confirm our understanding of systems level function. Examples of such models include cell cycle progression (Chen *et al.*, 2000), integrate and fire generation of heart pacemaker pulses (Zhang *et al.*, 2000) and cellular behavior in synchrony with the circadian cycle (Leloup and Goldbeter, 2003). A particular appeal of modeling is that models can be interrogated with *what if* type questions to improve our understanding of the system, or be used to make quantitative predictions in domains in which measurements are unavailable.

A central issue in developing computational models of biological systems is setting parameters such as rate constants of biochemical reactions, synthesis and decay rates of macromolecules, delays incurred in transcription of genes and translation of proteins, and sharpness of nonlinear effects (Hill coefficient) are examples of such parameters. Parameter values are usually determined by conducting *in vitro* experiments (*e.g.* (Wiedenmann *et al.*, 2004)). When parameter values are not available from experimental measurements, modelers often resort to hand-tuning during the model development process and publish the range of values of a parameter required to achieve a match between model output and observed data. In dynamical systems characterized by variations over time, concentrations of different molecular species (proteins, metabolites etc.) may also be of interest. In this setting, we encounter two difficulties. First, parameters measured by in vitro experiments may not be good reflections of the *in vivo* reality. And, second, some parameters in a system may not be amenable to experimental measurements. These limitations motivate the need to infer parameters in a computational model based on input-output observations and recent literature on computational and systems biology has seen intense activity along these lines (Liu and Niranjan, 2012).

One way of setting parameters systematically is based in techniques for search and optimization. For example, Mendes and Kell (1998) compared several optimization based algorithms for estimating parameters along biochemical pathways, concluding that no single approach significantly outperforms other available approaches. An alternate approach is the use of probabilistic Bayesian formulations to quantify uncertainties in the process of estimating parameters. Work described in Lawrence *et al.* (2006); Jayawardhana *et al.* (2008) fall into this category. A

to whom correspondence should be addressed

particular approach of interest is the method of Approximate Bayesian Computation (ABC) or likelihood-free inference. While this approach has its roots in population genetics (Tavaré *et al.*, 1997), where the likelihood is usually too complicated to write down in a computable form, it has attracted interest as a viable tool in systems biology parameter estimation problems ((Toni *et al.*, 2009; Secrier *et al.*, 2009)). In brief, the ABC approach assumes it is easier to simulate data from a model of interest than it is to compute and work with its likelihood under some assumed noise model. Hence a structured search could be carried out, in which one repeatedly simulates with parameter values sampled from a prior distribution, computes the error between simulated and observed data and decides to retain or reject a sample. At the search, all retained sample values define a posterior density in the space of parameters to be estimated.

2 METHODS

A new ABC based approach is developed by exploiting the fact that the values of the sloppy parameters can vary in a reasonable range, while the stiff parameters are determinants for evaluating the behavioral response and are therefore required to be precisely assigned. This method can be seen as a selective allocation of the computing budget for sloppy and stiff parameters. In the first step of proposed method, parameters of system are categorized into sloppy and stiff. Subsequently, all parameters are simultaneously estimated using a coarse acceptance criterion, and the value of sloppy parameters are set as the mean of posteriors. In the final step, the stiff parameters are re-estimated by considering tighter error tolerance. Consequently, this favorable allocation of computation budget alleviates the manual tuning for balancing accuracy and efficiency. Our approach is shown in Fig. 1.

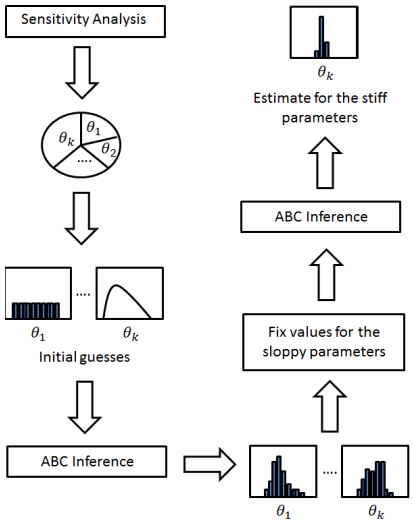


Fig. 1. Computational steps in the proposed approach: Starting from an initial distribution of parameter values, we carry out a coarse approximate Bayesian Computation (ABC) estimation of parameters. Following this, using sensitivity analysis we identify sloppy and stiff parameters of the system. The sloppy parameters are fixed to values determined by the coarse analysis. In the final stage, we estimate the stiff parameters of the system by running the ABC method to tighter error tolerance. This achieves a selective partitioning of the computational budget, and reliable estimates can be achieved within reasonable times.

2.1 ABC-SMC

The basic idea in ABC algorithms is to sample the unknown from a prior distribution, , synthesize data from the model under study, **X** **x** , where **x** is the initial condition and is the model, and accept as a sample for the posterior if the synthesized data **X** is close enough in some sense to the observations **X**. In its earliest form (Tavaré *et al.*, 1997), the generated particle was accepted only if **X** was identical to the observations **X**. It became immediately evident that this is an inefficient procedure because thousands of trails needed to be performed before accepting one of the generated particles. A modification to the scheme, introduced by Pitt and Shephard (1999) was to define a tolerance and accept particles when the discrepancy between **X** and **X** was within this.

Sequential Monte Carlo (SMC) method has been recently merged into the ABC framework (Sisson *et al.*, 2007; Toni *et al.*, 2009; Beaumont *et al.*, 2009), in which the acceptance criterion is extended as a sequence of tolerances . In general, SMC-based ABC methods draw the particles from the previous population by considering their weights, and perturb those particles around the space using the transition kernel, . The pseudo-observations are synthesized from the underlying model, **X** **x** , where **x** is the initial condition and is the dynamics. Particle is accepted and weighted if the discrepancy between synthetic data **X** and true dataset **X** is lower than the current tolerance .

In the sense of algorithmic setting, the tolerance schedule plays as the determinant in inference. Consequently, we adopts the innovative SMC-based ABC method (Del Moral *et al.*, 2012) which is capable to adaptively determine the tolerance schedule. The idea of this automatic scheme is to determine an appropriate reduction of the tolerance level based on the proportion of particles surviving under the current tolerance. If a large amount of particles remain 'alive', it implies the acceptance criterion is relatively loose and it is safe to make a jump for the next tolerance level. In contrast, if the ratio of 'alive' particles is low, this means particles are less likely to describe the posterior, therefore, a tiny movement should be considered.

2.2 Extended Fourier Amplitude Sensitivity Test

The *extended Fourier amplitude sensitivity test* (eFAST) (Saltelli *et al.*, 1999), is one of the popular sensitivity analysis techniques based on variance decomposition, being applicable for the nonlinear and non-monotonic systems.

The algorithm initially partitions the total variance of the dataset, evaluating what fraction of the variance can be determined by variations in the parameter of interest. This quantity, known as the sensitivity index, is calculated as

$$\mathbf{S} \quad \frac{\mathrm{Var}\quad \mathrm{E}\ \mathbf{X}}{\mathrm{Var}\ \mathbf{X}} \quad — \qquad (1)$$

'Translating' this definition into eFAST, the sensitivity is assessed by picking the samples for the parameter of interest with the highest frequency $_{\max}$, while the samples for the rest of the parameters are selected with the complementary frequencies . This process is repeated until the samples of each parameter is drawn with highest frequency once. An illustrative example of this cycling process is shown in Fig. 2. By using this sampling strategy and associating

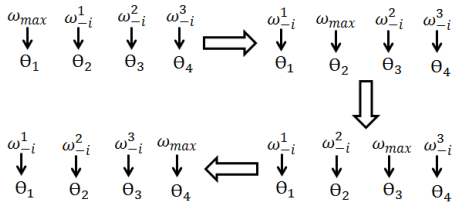


Fig. 2. When we wise to evaluate the sensitivity of parameter , its samples are drawn with the highest frequency $_{\max}$, while the samples for other parameters in the system are picked using the complementary frequencies . Through this process, all parameters in system should be assigned to the highest frequency once.

with Parseval's theorem, eFAST is capable to apportion the total variance (term in Eqn. 1) into the partial variance (term in Eqn. 1) caused by individual variation of the parameter, via the sum of Fourier coefficients. Algorithmically, the parameter sensitivity with respect to a specific state is evaluated by a fraction, given as Eqn. 1, where the numerator is the variance of outputs of the specific state. More specifically, the outputs adopted for numerator are synthesized by the parameter samples which are drawn from the frequency vector by setting the underlying parameter to the highest frequency. The denominator of this fraction is the summation of output variances of the same state, and these outputs are generated by different parameter samples, which are drawn from all possible combinations of frequency settings. For example, if we need to assess the sensitivity of parameter with respect to the second state **x** of a system that has four parameters, it can be calculated as

$$\mathbf{S} \quad \text{———————} \qquad (2)$$

where the subscript of shows which state in system is under study, and the superscript implies which parameter is sampled using the high frequency (i.e. the parameter of interest).

3 RESULTS

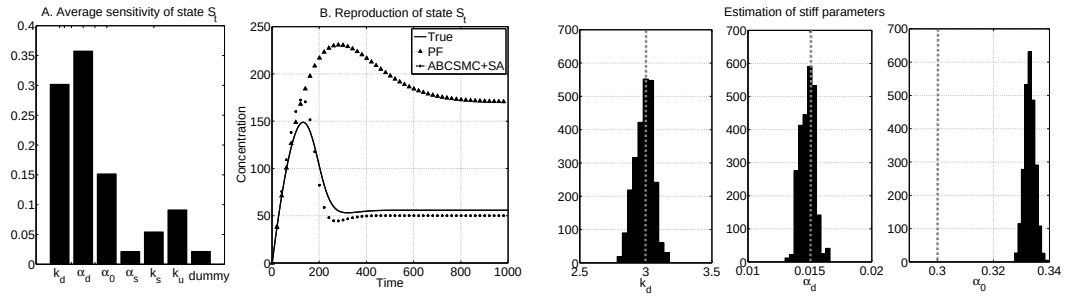
In the previous work (Liu and Niranjan, 2012), we considered the heat shock protein response system as a biological example for demonstrating the effectivenesses of Kalman algorithms and particle filter. When all parameters were unknown, being the hardest case considered, the non-parametric PF is able to recover four unknowns of six parameters. In order to discriminate the abilities of the proposed method and PF, a comparative study is carried out on the heat shock response system under the assumption that all parameters are unknown.

Fig. 3(a).A describes the average sensitivity of parameters with respect to state in the system, in addition, the sensitivity can be specifically evaluated at each time instant. Only the average sensitivity result is utilized in this work, but the decomposition of sensitivity is sometimes useful, e.g. to catalyze the specific reaction for achieving a rapid growth of species at a particular phase.

As shown in graph, the parameters , and are sensitive for producing system outputs, which are thus required to precisely infer. Making use of identical algorithmic settings (Liu and Niranjan, 2012), the estimation of stiff parameters from the proposed method are shown in Fig. 3(b) and the results of sloppy parameters are proposed in somewhere else. It can be easily seen that the particles of the stiff parameters center around the true values (is inferred with relatively low precision and larger variance, this is due to its less significance in comparison to the other two stiff parameters), whereas it fails to recover the true values of the sloppy parameters. Additionally, in the previously studied example, When the state is hidden in observations, PF was found to be incapable of precisely inferring the parameters and . Given the sensitivities of parameters with respect to the state , shown in Fig. 3(a).A, this is expected because these imprecisely inferred parameters govern the behavior of . Consequently, if is hidden in the observations, its corresponding stiff parameters are impossible to estimate.

In particular, assuming all parameters unknown, PF (four of six) seemingly wins the battle over ABC-SMC+SA (three of six) in terms of successful inferences. Fig. 3(a).B suggests that the proposed method slightly outperforms PF in terms of re-creating system dynamics, especially state . The more precise system re-characterization is also observed in other two states of system.

The effectiveness of proposed method has also been tested on repressilator, delay-driven oscillatory and cell cycle models, from the results of which the proposed method is capable for producing the promising inference within the affordable computational expense. In addition, in the cell cycle example, the proposed method can produce the relatively precise inference of parameters, whereas the standalone ABC-SMC cannot even terminate computation.



(a) Sensitivity analysis and state reproduction for heat shock system

(b) Stiff parameters of heat shock system from ABC+SA

Fig. 3. Sensitivity analysis, inference of parameters and system re-characterization of heat shock model. (a) A: Average sensitivity of parameters with respect to state . B: Reproduction of state by using true values, estimates from ABC+SA (dotted line) and PF (dotted triangle) respectively. (b) Histograms for the stiff parameters , and . The grey lines indicate the true values of parameters proposed in the literature.

4 CONCLUSION

In this work, we proposed an inference method for analyzing Systems Biology models that couples sensitivity analysis and approximate Bayesian computation. Our proposed method is particularly advantageous in difficult settings of estimating all (or most of) the parameters of a model from noisy observations, because it strikes a balance between accuracy and efficiency. The method exploits the fact that all parameters in model have different significance in characterizing model dynamics in terms of their sensitivities. By re-synthesis data from models with estimated parameters, we show that the values of parameters that are more critical (stiff parameters) need to be determined with care, while the sloppy parameters need not be estimated to high precision. To facilitate such inference, we have proposed a three stage strategy in which a selective computational budget allocation is implemented via sensitivity analysis, in which the sloppy group is estimated by a coarse search followed the re-estimation of stiff parameters to tighter error tolerances.

We have demonstrated the effectiveness of the proposed approach on three systems of oscillatory behavior and one of transient response. The results show that the introduction of favorable inference strategy allows to reduce the dimension of unknown parameters, and paves a potential way to tackle the complex problems. Additionally, the used ABC-SMC has attracted much interest due to its adaptivity in determining tolerance schedule and the 'no rejection' of particles allows to boost the efficiency via parallel computing. In the simple problem, e.g. the delay-driven oscillatory system, with performing similarly in accuracy, the proposed inference method expends much less computational cost than the existed ABC methods.

REFERENCES

Beaumont, M., Robert, C. P., Marin, J. M., and Cornuet, J. M. (2009). Adaptivity for abc algorithms: the abc-pmc scheme. *Biometrika*, **94**(4), 983–990.

Chen, K., Csikasz-Nagy, A., Gyorffy, B., Val, J., Novak, B., and Tyson, J. (2000). Kinetic analysis of a molecular model of the budding yeast cell cycle. *Mol. Biol. Cell*, pages 369–391.

Del Moral, P., Doucet, A., and Jasra, A. (2012). An adaptive sequential Monte Carlo method for approximate bayesian computation. *Stat. Comput*, **22**(5), 1009–1020.

Jayawardhana, B., Kell, D. B., and Rattray, M. (2008). Bayesian inference of the sites of perturbations in metabolic pathways via Markov chain Monte Carlo. *Bioinformatics*, **24**(9), 1191–1197.

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

Lawrence, N. D., Sanguinetti, G., and Rattray, M. (2006). Modelling transcriptional regulation using Gaussian processes. In *Advances in Neural Information Processing Systems 19*, page 785, Vancouver. MIT press.

Leloup, J.-C. and Goldbeter, A. (2003). Toward a detailed computational model for the mammalian circadian clock. *Proceedings of the National Academy of Sciences*, **100**(12), 7051–7056.

Liu, X. and Niranjan, M. (2012). State and parameter estimation of the heat shock response system using Kalman and particle filters. *Bioinformatics*, **28**(11), 1501–1507.

Mendes, P. and Kell, D. B. (1998). Non-linear optimization of biochemical pathways: applications to metabolic engineering and parameter estimation. *Bioinformatics*, **14**(10), 869–883.

Pitt, M. K. and Shephard, N. (1999). Filtering via simulation: Auxiliary particle filters. *J.Amer.Statist.Assoc*, **94**(446), 590–599.

Saltelli, A., Tarantola, S., and Chan, K. P.-S. (1999). A Quantitative Model-Independent Method for Global Sensitivity Analysis of Model Output. *Technometrics*, **41**(1), 39–56.

Secrier, M., Toni, T., and Stumpf, M. P. H. (2009). The ABC of reverse engineering biological signalling systems. *Mol. BioSyst.*, **5**, 1925–1935.

Sisson, S. A., Fan, Y., and Tanaka, M. M. (2007). Sequential Monte Carlo without likelihoods. *Proc. Natl. Acad. Sci.*, **104**, 1760–1765.

Tavaré, S., Balding, D. J., Griffiths, R. C., and Donnelly, P. (1997). Inferring coalescence times from dna sequence data. *Genetics*, **145**, 505–518.

Toni, T., Wlech, D., Strelkowa, N., Ipsen, A., and Stumpf, M. P. H. (2009). Approximate Bayesian computation scheme for parameter inference and model seletion in dynamical systems. *J. R. Soc*, **6**, 187–202.

Wiedenmann, J., Ivanchenko, S., Oswald, F., Schmitt, F., Röcker, C., Salih, A., Spindler, K.-D., and Nienhaus, G. U. (2004). EosFP, a fluorescent marker protein with UV-inducible green-to-ren fluorescence conversion. *PNAS*, **101**, 15905–15910.

Zhang, H., Holden, A. V., Kodama, I., Honjo, H., Lei1, M., Varghese, T., and Boyett, M. (2000). Mathematical models of action potentials in the periphery and center of the rabbit sinoatrial node. *American Journal of Physiology - Heart and Circulatory Physiology*, **279**(1), 397–421.

M^{3}D: a kernel-based test for shape changes in methylation profiles

Tom Mayo [1], Gabriele Schweikert [1,2] and Guido Sanguinetti [1]

[1]IANC, School of Informatics, University of Edinburgh
[2]Wellcome Trust Centre for Cell Biology, University of Edinburgh

I. Introduction

DNA methylation is an epigenetic mark associated with gene transcription and imprinting, retrotransposon silencing and cell differentiation [1]. Methylation occurs when a methyl group is attached to a cytosine and this is predominantly observed in the CpG context in mammals. Classically, methylation of CpG islands (CGIs) in promoter regions is associated with gene silencing, however, recent studies have shown that CpG methylation correlates with gene expression in a more complex, context-dependent manner [2]. Moreover, the shape of the methylation profile is also an important factor in predicting gene expression [3].

Bisulfite treatment of DNA followed by next generation sequencing provides quantitative methylation data with base pair resolution. Unmethylated cytosines are deaminated into uracils, which amplify as thymines during PCR. Reads are then aligned to a reference genome, permitting changes of C to T. The resulting counts of cytosine and thymine at registered cytosine loci form the basis of further analysis. This general procedure has been adapted in various ways, with reduced representation bisulfite sequencing (RRBS) being the most widely used. RRBS involves using a restriction enzyme such as MspI (or TaqI) to cleave the DNA at CCGG (or TCGA) loci and selecting short reads for sequencing, resulting in greater coverage of CpG dense regions at lower cost.

Several methods have been proposed to identify differentially methylated regions (DMRs). Early methylation studies used Fisher's exact test to identify differentially methylated cytosines (DMCs), which were then chained together to call DMRs. Later methods have continued to call DMCs and chain them into DMRs, focusing on improving the initial step. BSmooth, the most widely used of the later methods, uses local likelihood smoothing to generate methylation profiles for each sample and calls DMCs based on the profile values [4]. Other methods, such as BiSeq [5], call DMCs based on a beta-binomial model of methylation, in this case using the Wald test to improve statistical power. To our knowledge, there are no methods that test higher order properties of the methylation profiles across regions, such as shape.

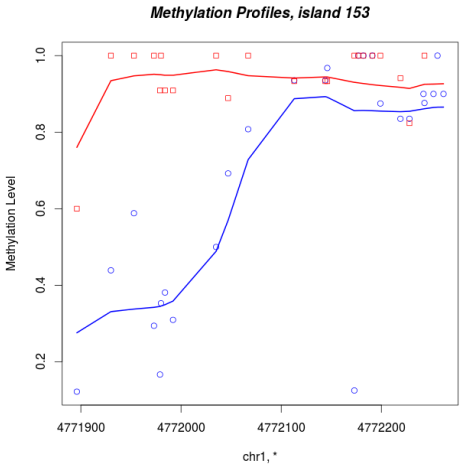


Figure 1: A DMR identified uniquely by M^{3}D. The profile shape changes in the first 100-200bp. Blue is from healthy breast cells, red from mammary cancer cells.

Here we present the M^{3}D method, a non-parametric test for identifying DMRs and, to our knowledge, the first to explicitly account for shape. Our method is based on the maximum mean discrepancy (MMD), a recent technique from the machine learning literature which tests whether two samples have been generated from the same probability distribution [6, 7]. This method has already been applied to ChIP-Seq data [8]. Here, we adapt the method for the specific challenges of RRBS data, namely that the sampling frequency (the coverage) of an in-

dividual cytosine is unrelated to its methylation level. We demonstrate the performance of M^{3}D against existing methods on real and simulated data. An example DMR uniquely identified by M^{3}D is shown in Figure 1.

II. Methods

The M^{3}D method is designed to analyse aligned methylation data. Rather than testing individual cytosines and pooling them into putative DMRs, M^{3}D considers changes in the methylation profile's shape over a given genomic region. The maximum mean discrepancy is calculated over each region and adjusted to account for changes in the coverage profile across samples. Finally, we use a data-driven approach to compare test statistics based on the empirical likelihood of observing between-group differences among replicates.

Regions can either be pre-defined, such as a list of promoter regions, or generated from the data by clustering the observed CpGs. Here, we adopt the method used in [5] to define clusters of CpGs. We restrict our analysis to CpGs only and combine data from both strands.

I. Maximum Mean Discrepancy

Formally, the MMD is defined as follows. Let $\mathcal{F}$ be a class of functions $f : \mathcal{X} \to \mathbb{R}$ over a metric space $\mathcal{X}$ with Borel probability measures p, q. We define the MMD as:

$$MMD[\mathcal{F}, p, q] = \sup_{f \in \mathcal{F}} \left(\mathbf{E_p}[f(x)] - \mathbf{E_q}[f(x)] \right) \tag{1}$$

Intuitively, we are finding the mean over a bounded function that maximises the difference between the probability distributions. For a sufficiently dense function class, this is equal to 0 if, and only if, $p = q$. Choosing $\mathcal{F}$ to be the unit ball in a reproducing kernel Hilbert space (RKHS) on $\mathcal{X}$ provides a searchable class of functions that retains this result [6].

In practice, for $X = \{x_1, ..., x_m\}, Y = \{y_1, ..., y_n\}$ observations independently and identically distributed (i.i.d.) from p and q respectively, we can approximate the MMD metric with an appropriate choice of kernel, k, giving rise to a feature representation in the RKHS:

$$MMD[X, Y] = \left[\frac{K_{xx}}{m^2} - \frac{2K_{xy}}{mn} + \frac{K_{yy}}{n^2} \right]^{\frac{1}{2}} \tag{2}$$

Where $K_{xx} = \sum_{i,j=1}^{m} k(x_i, x_j)$, $K_{xy} = \sum_{i,j=1}^{m,n} k(x_i, y_j)$, and $K_{yy} = \sum_{i,j=1}^{n} k(y_i, y_j)$

II. The M^{3}D statistic

We represent a RRBS data set as a set of vectors x_i, where each x_i is composed of the genomic location of a cytosine C_i, and the methylation status of that C_i on one mapped read, $x_i = (C_i, Meth_i)$. Thus, there are as many x_is in a data set as the number of mapped cytosines (within a CpG context). In order to define an MMD between data sets, we need to define a kernel function operating on pairs of vectors x_i, x_j in order to evaluate equation (2). A natural choice is a composite kernel given by the product of a radial basis function (RBF) kernel on the genomic location and a string kernel on the methylation status: $k_{STR}(x_i, x_j) = 1$ if $Meth_i = Meth_j$, 0 otherwise. The RBF kernel takes the form:

$$k_{RBF}(x_i, x_j) = exp[-(C_i - C_j)^2 / 2\sigma^2]$$

retaining spatial information at a scale determined by the hyper-parameter σ, which corresponds to the distance along the genome that displays methylation correlation. We model this parameter independently for each region, R, to reflect the local correlation structure, as $\sigma_R^2 = \bar{x}^2/2$, for $x \in R$, a heuristic suggested in [7]. Here $\bar{x}$ refers to the median distance of all observations in region R across the data sets being compared. MMD distances computed using the above procedure would capture both differences in coverage profiles and differences in methylation profiles. A particular challenge of bisulfite sequencing data, and a central tenet of the RRBS procedure (9), is that the frequency with which a cytosine site is tested (the coverage) is unrelated to the methylation status. This poses a challenge in all bisulfite sequencing analysis, as the sampling distribution becomes a confounding factor in our attempt to understand methylation. We control for changes in the coverage profile by subtracting the analogous MMD of the coverage; the M^{3}D metric is then given by:

$$M^3D[X,Y] = MMD[X,Y,k_{full}] - MMD[X,Y,k_{RBF}] \quad (3)$$

where $k_{full}(x_i, x_j) = k_{STR}(x_i, x_j)k_{RBF}(x_i, x_j)$ as described above and the MMD terms are as in equation (2).

The last term in equation (3) represents the MMD of the data on a methylation-blind subspace. This implies that, in the large sample limit when the sample estimate of the MMD converges to the exact MMD of equation (1), the M^{3}D statistic is non-negative.

The M^{3}D statistic will therefore be different from zero when there is a change in the methylation profile, independently of a change in the coverage profile. As a consequence, M^{3}D between replicate RRBS experiments (which do not necessarily have identical coverage) should be zero or, equivalently, the full MMD should be equal to the methylation-blind MMD. This is borne out in the data; the metrics strongly agree over replicates. Testing equality of metrics over 102 ENCODE RRBS data sets gives an R^2 of 0.95.

We use the M^{3}D as a test-statistic. *P*-values for testing regions are defined as the empirical probability of observing the mean value of the cross-group statistics among the replicates within groups. For a given region, this is calculated by finding the mean, μ of the M^{3}D statistics in all the cross group comparisons of that region. The associated p-value is then the proportion of all of the inter-replicate M^{3}D statistics, over all regions and samples, that are greater than or equal to μ. We use the Benjamini-Hochberg procedure to calculate false discovery rates (FDRs), rejecting clusters at a 1% significance level. Since each test corresponds to an entire region, this correction is less punitive than methods testing each cytosine location.

III. Simulations

Since almost all of the ENCODE RRBS data provides 2 replicates, and existing methods lose power with fewer replicates, we ran simulations to test the M^{3}D method's performance with 2 replicates in each group. We chose cell line K562 (GEO: GSM683780), a human leukemia line, as our control group. To simulate a realistically different coverage profile, we took the coverage profile from another cell line, MCF7, a breast cancer cell line (GEO: GSM683787). Both data sets are publicly available from the ENCODE consortium [10]. We grouped the CpGs into clusters and picked the first 1000 on chromosome 1. To simulate clusters that didn't change, we sampled methylation reads for locus i in replicate j from a binomial distribution $X_{i,j} \sim B(n_{i,j}, p_{i,j})$ where $n_{i,j}$ is the coverage and $p_{i,j}$ is determined from the methylation profile of the control group. We chose to alter the profiles of 250 randomly selected clusters, by determining a short region and altering the $p_{i,j}$ values, hyper- or hypo-methylating them according to their control levels. For this experiment, we chose to alter a region that was at least 100bp long and containing at least 4 CpGs. To avoid skewing the results when computing p-values, we compared the test-statistic to the replicates of the real data set only. To test robustness against lower coverage, we reduced the coverage at all sites by 75% and 50% and resampled the methylation counts according to a binomial distribution with the probability of methylation being the observed methylation ratio at the site. To compare our results against other publicly available methods, we used BSmooth [4].

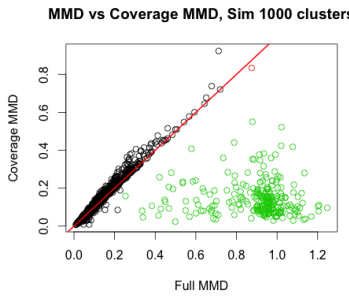
IV. Breast Cancer Data

To test the M^{3}D method on real data, we compared two tracks from the ENCODE consortium. RRBS data from normal breast cells (GEO: GSM683834) were compared against mammary cancer cells (MCF-7, GEO: GSM683787). In both cases, data was produced, pre-processed and aligned to the hg19 genome by the Myers Lab at the HudsonAlpha Institute for Biotechnology. Again, BSmooth was applied for comparison.

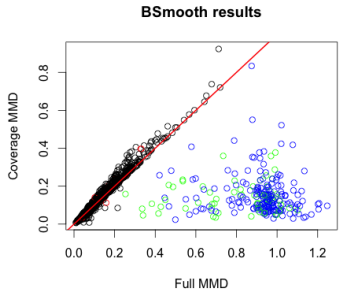
III. Results

I. Simulations

In our simulation, M^{3}D correctly identified 249 of the 250 ground truth DMRs, while BSmooth identified 67. Additionally, BSmooth had 19 type 1 errors, while M^{3}D had none. The M^{3}D method remained robust to resampling the data at 75% and 50% coverage. The test-statistics are plotted in Figure 2a, where we see that the M^{3}D framework identifies a clear difference between the DMRs and the unchanged islands that BSmooth does not accurately detect (Figure 2b).

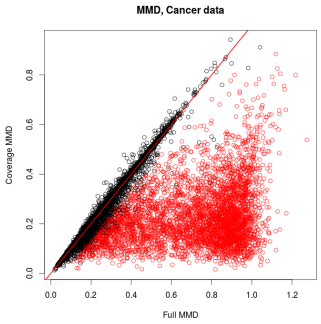


(a) M^3D

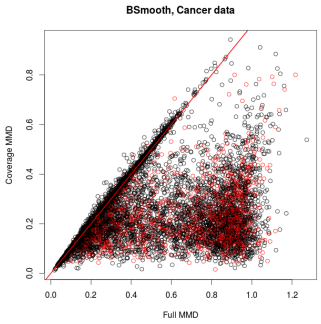


(b) BSmooth

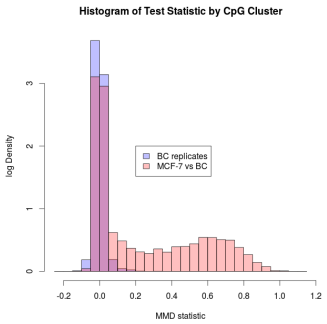
Figure 2: Simulation Results. Plots show the methylation-blind metric against the full MMD. The test statistic is their difference. Each point is a CpG cluster. Black are unchanged, Green are correctly called DMRs, Red and Blue are type 1 and 2 errors. (a) M^3D identifies a clear relationship and calls almost all of the clusters. (b) BSmooth misses 183 clusters.



(a) M^3D



(b) BSmooth



(c) Histogram of Test-Statistic

Figure 3: MCF7 vs BC Breast Cells. Black dots are uncalled clusters, red are called. (a) Between-group clusters as called by M^3D. (b) BSmooth identifies far fewer. (c) Histogram of test-statistic, log-density scale. The empirical distribution of replicate statistics is blue. The red bulge to the right represents DMRs in (a).

II. Breast Cancer Data

We identified 15,458 CpG clusters from the data. Of these, BSmooth identified 1425 that contained DMRs, while M^3D identified 4,072. The methods agree on 1121 of the clusters. Figures 3a and 3b show the results of between-group testing by M^3D and BSmooth respectively. There is a striking similarity to Figure 2a, indicating that these changes are not to be expected from biological variation. With BSmooth, many of these differences are missed (Figure 3a). The empirical test-statistic distribution is plotted with the cross-group test-statistics in Figure 3c.

IV. Discussion

We proposed the first kernel-based test for DMRs and have demonstrated significant advances over an existing, widely used method, BSmooth [4]. The M^3D framework was able to detect 249 out of 250 simulated DMRs without falsely rejecting any clusters and identified 2951 extra DMRs in a real world data set.

The benefits of the M^3D approach, which we have outlined above, are due to a number of factors. Sites of lower coverage are not dominated

by their neighbours in a smoothing process, nor are they thrown out. Correlations among neighbouring cytosines are measured on a cluster-by-cluster basis and adapted to each region. Finally, the test-statistic is measured empirically against the variability of the replicates. This provides more confidence that the methylation changes of identified DMRs are of a magnitude that cannot be explained by inter-replicate variability.

The M^{3}D framework was developed with RRBS data in mind. Given its robustness to lower coverage, we believe it may also be suited to whole genome bisulfite sequencing (WGBS) data.

Acknowledgements

Funding: TM is funded by grants from the UK EPSRC, BBSRC and MRC. GSch currently holds a Marie Curie Fellowship. GS is funded by European Research Council through grant MLCS306999.

References

[1] Louise Laurent, Eleanor Wong, Guoliang Li, Tien Huynh, Aristotelis Tsirigos, Chin Thing Ong, Hwee Meng Low, Ken Wing Kin Sung, Isidore Rigoutsos, Jeanne Loring, et al. Dynamic changes in the human methylome during differentiation. *Genome research*, 20(3):320–331, 2010.

[2] Katherine E Varley, Jason Gertz, Kevin M Bowling, Stephanie L Parker, Timothy E Reddy, Florencia Pauli-Behn, Marie K Cross, Brian A Williams, John A Stamatoyannopoulos, Gregory E Crawford, et al. Dynamic DNA methylation across diverse human cell lines and tissues. *Genome research*, 23(3):555–567, 2013.

[3] Nathan D VanderKraats, Jeffrey F Hiken, Keith F Decker, and John R Edwards. Discovering high-resolution patterns of differential DNA methylation that correlate with gene expression changes. *Nucleic acids research*, 41(14):6816–6827, 2013.

[4] Kasper D Hansen, Benjamin Langmead, Rafael A Irizarry, et al. BSmooth: from whole genome bisulfite sequencing reads to differentially methylated regions. *Genome Biol*, 13(10):R83, 2012.

[5] Katja Hebestreit, Martin Dugas, and Hans-Ulrich Klein. Detection of significantly differentially methylated regions in targeted bisulfite sequencing data. *Bioinformatics*, 29(13):1647–1653, 2013.

[6] Arthur Gretton, Karsten M Borgwardt, Malte Rasch, Bernhard Schölkopf, and Alexander J Smola. A kernel method for the two-sample-problem. *Advances in neural information processing systems*, 19:513, 2007.

[7] Arthur Gretton, Karsten M Borgwardt, Malte J Rasch, Bernhard Schölkopf, and Alexander Smola. A kernel two-sample test. *The Journal of Machine Learning Research*, 13:723–773, 2012.

[8] Gabriele Schweikert, Botond Cseke, Thomas Clouaire, Adrian Bird, and Guido Sanguinetti. MMDiff: quantitative testing for shape changes in ChIP-Seq data sets. *BMC genomics*, 14(1):826, 2013.

[9] Hongcang Gu, Zachary D Smith, Christoph Bock, Patrick Boyle, Andreas Gnirke, and Alexander Meissner. Preparation of reduced representation bisulfite sequencing libraries for genome-scale DNA methylation profiling. *Nature protocols*, 6(4):468–481, 2011.

[10] ENCODE Project Consortium et al. An integrated encyclopedia of DNA elements in the human genome. *Nature*, 489(7414):57–74, 2012.

Structuration of the bacterial replicon space

Olivier POIRION* [1] and Bénédicte LAFAY[1]

[1]Université de Lyon, CNRS UMR5005-Laboratoire Ampère, École Centrale de Lyon, 36 avenue Guy de Collongue, F-69134 Écully

In the context of studying the relationships among bacterial replicons, i.e., chromosomes, extra-chromosomal and essential replicons, and plasmids, we investigate the structuration of these genomic elements according to their replication, segregation and maintenance systems. Standard method fail to describe the complexity of the genetic events that occur in the evolution and adaptation of these elements. Given a set of genes of interest linked functionally to these genomic elements and used as variables, the organization of bacterial replicons was studied using clustering methodology and evaluation indexes. Results show a dual functional and taxonomical structuration of the replicon space. This led to results with strong biological implications. Indeed, we were able to characterize the third class of replicons relative to chromosomes and plasmids, and to propose novel defining criteria for these genomic elements. Beyond biological relevance, our study sets the basis for further analyses (workflow improvement/enrichment, classifications...) to bring to light on the driving forces of genome evolution.

1 Introduction

Bacterial genomes are constistuted of different types of replicons [10], separated into chromosomes and plasmids. The former are the essential component of the genome whereas the latter are dispensable to the host bacterium. Numerous inter- and intra-species DNA exchanges have been reported between chromosomes and plasmids [13]. Interactions and recombinations between these are expected to result in a complex set of gene homologies and thus in the blurring of genome and organismal evolution. Furthermore, some bacteria harbour in their genome replicons that exhibit both chromosomal and plasmidic features [12] and can be defined as **E**xtra-**C**hromosomal **E**ssential **R**eplicons (ECERs). Although some authors consider ECERs to be adapted plasmids [6], the origin, the genetic adaptations and the roles of these elements remain unclear. Moreover, despite being essential to the cell, no universal diagnostic feature has been discovered to date. A realistic hypothesis is that these elements possess adapted genetic **Re**plication, **S**egregation and **Ma**intenance **S**ystems (ReSMaS) to get synchronised with the cell cycle [17]. We thus investigated the discrimination of bacterial replicons according to their distributions of genes coding for proteins related to ResMaS. We first identified and assessed ResMaS proteins from all bacterial genomes available, and used them to build clusters of functional homologs (Section 2). We used these as attributes to describe the bacterial replicons. The formed dataset was then clusterized (Section 3) and visualized (Section 4).

2 Dataset construction

Proteins involved in the replication and segregation of the replicons and the cell cycle were used to build annotated clusters of functional homologs using BLAST [1] and TRIBE-MCL [2] clustering algorithm. A query dataset was constructed based on chosen proteins homolog families from ACLAME [11] and from KEGG [9] using KEGG BRITE hierarchy. The query set was then used as input in a BLAST analysis (10^{-5} E_{value} cutoff) to identify putative homologous proteins among all bacterial protein sequences available from the Genbank database [4] on 30/11/2012. An all-vs-all BLAST analysis was conducted and the resulting score matrix was used as input to TRIBE-MCL to build clusters of homologous proteins. Finally, a cleaning procedure was undertaken by comparing the Pfam functional do-

*olivier.poirion@ec-lyon.fr

mains [3] of the proteins in the obtained clusters to those of the proteins of the query dataset. Clusters with too distant domains distributions were removed. Our final dataset amounted to **6098** clusters and **4928** replicons. The numbers of chromosomes, plasmids and ECERs included to the analysis are **2016**, **2744** and **129**, respectively. The very high rate of functional homologs per cluster attested of the MCL outputs reliability.

Given R, the set of replicons and $Cl = \{C_1, ..., C_k\}$ the set of k clusters of proteins. For $r \in R$, a characteristic vector of r: v^r can be defined by:

$$v^r = (N^r_{C_1}, ..., N^r_{C_k}) \quad (1)$$

where $N^r_{C_i}$ is the number of proteins of r in C_i with $C_i \in Cl$. Let V^R be the set of v_r for all $r \in R$.

Because the representation of the bacterial replicons is biased according to the host species taxonomy, we also normed the data according to the host genus and the replicon type. Let $R^{\{type\}}_{taxa}$ be all the replicons from R of type *type*, with $type = \{chromosome, plasmid, ECER\}$ and from the genome of a bacterial species of the genus *taxa*. $v^{\{taxa,type\}}$ is then defined by:

$$v^{\{taxa,type\}} = \frac{1}{|R^{\{type\}}_{taxa}|}\left(\sum_{r \in R^{\{type\}}_{taxa}} N^r_{C_1}, ..., \sum_{r \in R^{\{type\}}_{taxa}} N^r_{C_k}\right) \quad (2)$$

Let $\bar{V}^R_{genus}$ the set of $v^{\{taxa,type\}}$ for all *taxa* of genera of represented species and for all *type*. Each dataset V^R or $\bar{V}^R_{genus}$ can be seen as a bipartite graph where the first set of nodes is either R or all the non-empty $R^{\{type\}}_{taxa}$, and the second set of nodes corresponds to the protein clusters. An edge exists between a protein cluster C_i and a replicon r (or $R^{\{type\}}_{taxa}$) if $v^r[i] \neq 0$. Its weight is then equal to $v^r[i]$.

3 Clustering of bacterial replicons

The objective of the clustering of the replicons is to characterize the hidden structure of bacterial replicon according to the replicon ResMaS: a grouping of close replicons will indicate that these replicons are linked from a both evolutionary and functional standpoints.

Two clustering procedures were applied to datasets V^R and $\bar{V}^R_{genus}$: i) the community detection algorithm: INFOMAP [15] on the bipartite graphs of the dataset, and ii) a dimension reduction procedure consisting of a principal component analysis followed by a hierarchical clustering using the WARD algorithm [16].

3.1 Evaluation methodology

Two criteria were mesured to assess the clustering solution reliability:

• An external index, the *V-measure*, was computed as the harmonic mean of two other criteria: *homogeneity* and *completeness* [14]. Homogeneity denotes how uniform clusters are towards a class of reference. The completeness indicates whether reference classes are embedded within clusters. These three indices vary between 0 and 1, the values closest to 1 reflecting the good quality of the clustering solution. The type of replicon (*i.e.*, plasmid, chromosome, or ECER), the taxonomic affiliation (phylum or class) of chromosomes and that of plasmids, were used alternatively as reference classes.

• Additionally, the stability criterion of individual clusters for a given clustering result [7] was evaluated. The original version of this index computes the mean of the Jaccard distance between each cluster for a given clustering solution and the re-sampling of the clustering results. This index varies between 0 and 1 with values closest to 1 indicating that the clustering solution is highly stable. Each cluster stability result was weighted by the size of the cluster.

3.2 Results

Obtained clusters are stables overall (Table 2) in the whole and thus are informative. Both methods succeed in separating the chromosomes from the plasmids, thereby underlining the difference between their respective ResMaS. For some of the ECER-containing genera (*Brucella, Burkholderia, Vibrio, Leptospira*), specific ECERs clusters are retrieved (Table 1), which **asserts the specificity of the REsMaS of these replicons**. Furthermore ECERs from some taxa seem to be closer either to chromosomes or plasmids (highlighted in blue or yellow in Table 1, respectively). **This leads us to conclude that different genetic mechanisms are at work in the ECER formation and regulation.** Finally, somewhat unexpectedly, INFOMAP structures the plasmids according to the taxonomy of their host, thus revealing a functional link between plasmid ResMaS and their host identity. This in turn is not observed with the PCA+WARD procedure which is mainly due to the *curse of dimensionality* problem [8]. This highlights the superiority of INFOMAP to cluster our high-dimensional datasets.

4 Visualisation of bacterial replicons

Replicons were further analysed using the **M**ulti-**D**imensional **S**caling algorithm [8] and the *cosine* distance to compute inter-replicon distance [5] (Figures). The visualisation of the spatial organisation of the replicons confirms the clustering results and, moreover shows that **ECERs are located primarily between chromosomes and plasmids**.

5 Conclusion

Through the characterisation of bacterial replicons using their ResMaS, we were able to identify and describe a novel class of genomic element, the ECERs, which were incompletely defined until now. Various degrees of proximity to chromosomes and plasmids seem to exist among the ECERs suggesting the existence of several type of ECERs.
These findings were made possible by investigating high-dimensional data using data-mining and machine learning methodologies. Additional analyses performed comprised regressions and classifications in order to i) better characterise the ResMas skewd distribution of the different types of replicons, and ii) identify yet undescribed ECERs among plasmids. When performing such analyses, the various sources of analytical bias encountered: false homologs detection, choice of algorithms and parameters, data sampling bias..., must be acknowledged and taken into account. Finally, each detected trend should ideally be interpretable by genomic or evolutionary mechanisms.

References

[1] Camacho, C., Coulouris, G., Avagyan, V., Ma, N., Papadopoulos, J., Bealer, K., and Madden, T. L. BLAST+: architecture and applications. *BMC Bioinformatics 10*, 1 (2009), 421.

[2] Enright, A. J., Van Dongen, S., and Ouzounis, C. A. An efficient algorithm for large-scale detection of protein families. *Nucleic Acids Research 30*, 7 (2002), 1575–1584.

[3] Finn, R. D., Bateman, A., Clements, J., Coggill, P., Eberhardt, R. Y., Eddy, S. R., Heger, A., Hetherington, K., Holm, L., Mistry, J., et al. Pfam: the protein families database. *Nucleic acids research 42*, D1 (2014), D222–D230.

[4] Geer, L. Y., Marchler-Bauer, A., Geer, R. C., Han, L., He, J., He, S., Liu, C., Shi, W., and Bryant, S. H. The NCBI BioSystems database. *Nucleic Acids Research 38*, Database issue (2010), D492–D496.

[5] Han, J., Kamber, M., and Pei, J. *Data mining: concepts and techniques*. Morgan kaufmann, 2012.

[6] Harrison, P., Lower, R., Kim, N., and Young, J. Introducing the bacterial chromid': not a chromosome, not a plasmid. *Trends in microbiology* (2010).

[7] Hennig, C. Cluster-wise assessment of cluster stability. *Computational Statistics & Data Analysis 52*, 1 (2007), 258–271.

[8] Izenman, A. Modern multivariate statistical techniques: regression. *Classification, and Manifold Learning Springer Texts in Statistics, New York* (2008).

[9] Kanehisa, M., Goto, S., Sato, Y., Furumichi, M., and Tanabe, M. KEGG for integration and interpretation of large-scale molecular data sets. *Nucleic acids research 40*, Database issue (Jan. 2012), D109–14.

[10] Lederberg, J. Plasmid (1952-1997). *Plasmid* (1998).

[11] Leplae, R., Lima-Mendez, G., and Toussaint, A. ACLAME: A CLAssification of Mobile genetic Elements, update 2010. *Nucleic Acids Research 38*, Database issue (2010), D57–D61.

[12] Mackenzie, C., Kaplan, S., and Choudhary, M. Multiple chromosomes. *Microbial evolution: gene establishment, survival, and exchange* (2004), 82–101.

[13] Passot, F. M., Calderon, V., Fichant, G., Lane, D., and Pasta, F. Centromere binding and evolution of chromosomal partition systems in the burkholderiales. *Journal of bacteriology 194*, 13 (July 2012), 3426–36.

[14] Rosenberg, A., and Hirschberg, J. V-Measure: A conditional entropy-based external cluster evaluation measure. *Computational Linguistics*, June (2007), 410–420.

[15] Rosvall, M., and Bergstrom, C. T. Maps of random walks on complex networks reveal community structure. *Proceedings of the National Academy of Sciences of the United States of America 105*, 4 (Jan. 2008), 1118–23.

[16] Ward Jr, J. H. Hierarchical grouping to optimize an objective function. *Journal of the American statistical association 58*, 301 (1963), 236–244.

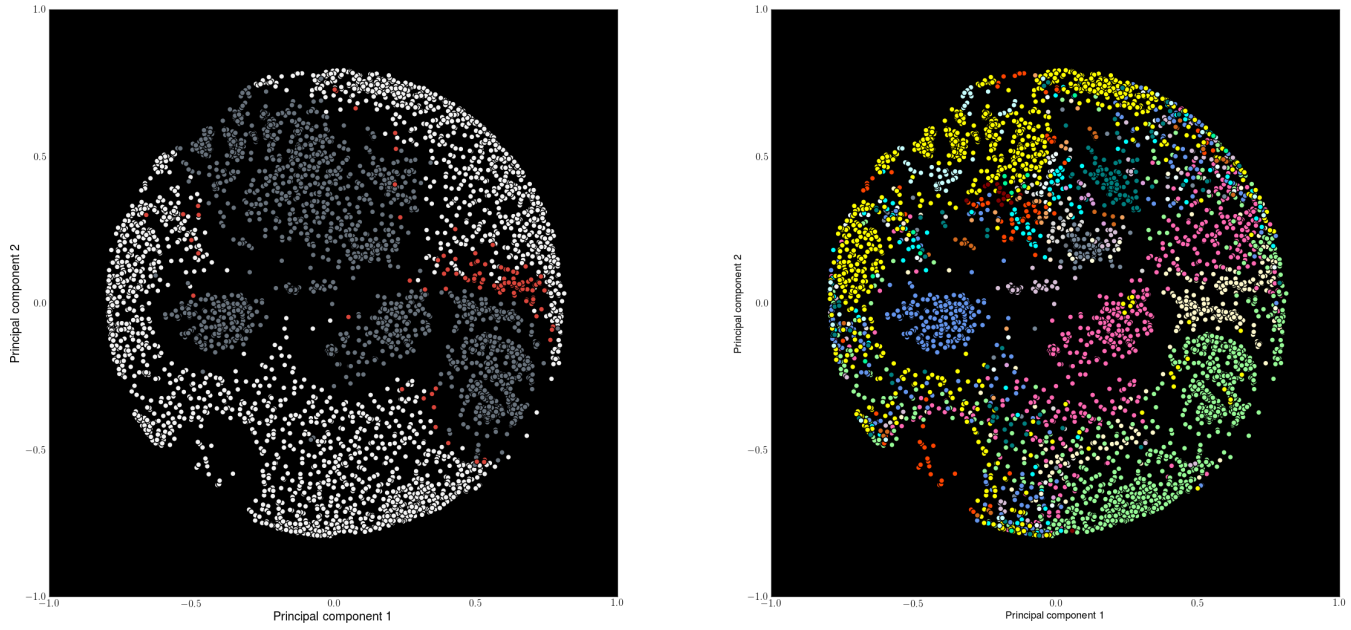
[17] Yamaichi, Y., Gerding, M. A., Davis, B. M., and Waldor, M. K. Regulatory cross-talk links vibrio cholerae chromosome ii replication and segregation. *PLoS genetics 7*, 7 (2011), e1002189.

Table 1: *Characteristic of ECER-containing clusters according to genus. **C** is the number of ECER-containing clusters for a given genus. %chr, %pl and %ECER are the percentage of chromosomes, plasmid, and ECER, respectively, in the ECER-containing clusters and $E(\triangle^C)$ is the stability of the ECER-containing clusters.*

Genus	C	$\%_{chr}$	$\%_{pl}$	$\%_{ECERs}$	$E(\triangle^C)$
Agrobacterium	3	0.27	0.38	0.35	0.4
Aliivibrio	1	0.0	0.0	1.0	0.95
Anabaena	1	0.96	0.02	0.02	0.9
Asticcacaulis	1	0.96	0.03	0.01	0.97
Brucella	1	0.0	0.05	0.95	0.87
Burkholderia	2	0.64	0.19	0.17	0.73
Butyrivibrio	1	0.0	0.5	0.5	0.83
Chloracidobacter	1	0.91	0.08	0.01	0.86
Cupriavidus	1	0.73	0.09	0.18	0.72
Cyanothece	1	0.0	0.94	0.06	0.61
Deinococcus	1	0.0	0.96	0.04	0.61
Ilyobacter	1	0.91	0.08	0.01	0.86
Leptospira	1	0.0	0.13	0.88	1.0
Nocardiopsis	1	0.91	0.09	0.0	0.97
Ochrobactrum	1	0.0	0.05	0.95	0.87
Paracoccus	1	0.96	0.03	0.01	0.97
Photobacterium	1	0.96	0.03	0.01	0.79
Prevotella	1	0.95	0.02	0.02	0.92
Pseudoalteromonas	1	0.96	0.03	0.01	0.79
Ralstonia	1	0.73	0.09	0.18	0.72
Rhodobacter	1	0.0	0.6	0.4	0.71
Sinorhizobium	2	0.0	0.87	0.13	0.87
Sphaerobacter	1	0.0	0.5	0.5	1.0
Sphingobium	2	0.7	0.29	0.01	0.94
Thermobaculum	1	0.91	0.08	0.01	0.86
Variovorax	1	0.73	0.09	0.18	0.72
Vibrio	1	0.0	0.0	1.0	0.95

Table 2: *V-measure and stability scores obtained for the clustering of V^R and $\bar{V}^R_{genus}$. k is the input number of clusters for WARD and pc is the number of retained PCA components.*

	Indice	PCA+WARD		INFOMAP	
data		V^R	$\bar{V}^R_{genus}$	V^R	$\bar{V}^R_{genus}$
Parameters		k:200 pc:30	k:100 pc:15	iter:500	
Clusters number		175	75	223	77
Explained variance		0.57%	0.58%		
Stability $\triangle^{Kl}$		0.85	0.74	0.82	0.76
Replicon type	*homogeneity*	0.93	0.83	0.80	0.63
	completeness	0.25	0.20	0.15	0.15
	V-measure	0.43	0.32	0.25	0.24
Chromosome Phylum	*homogeneity*	0.93	0.80	0.93	0.69
	completeness	0.35	0.40	0.60	0.61
	V-measure	0.51	0.53	0.73	0.65
Chromosome class	*homogeneity*	0.93	0.80	0.85	0.64
	completeness	0.16	0.58	0.80	0.82
	V-measure	0.66	0.67	0.82	0.72
Plasmid phylum	*homogeneity*	0.06	0.01	0.88	0.78
	completeness	0.16	0.14	0.33	0.35
	V-measure	0.08	0.02	0.48	0.48
Plasmid class	*homogeneity*	0.07	0.02	0.84	0.74
	completeness	0.28	0.36	0.43	0.51
	V-measure	0.12	0.03	0.57	0.60



(A) Replicons projection per type

(B) Replicons projection per taxonomical group

Figure 1: *Visualisation of the MDS Projection of V^R according to* cosine *distance. (A) Projection per type: chromosome (grey), plasmid (white) and ECERs (red). (B) Projection according to taxonomical class: Alpha- (pink), Beta- (light yellow), Gamma- (green), Delta- (grey), Epsilon- (light purple) Proteobacteria; and to phyla: Deinococcus-Thermus (white), Actinobacteria (blue), Cyanobacteria (cyan), Acidobacteria (beige), Spirochaetes (red), Firmicutes (yellow), Chlamydiae (brown), Bacteroidetes (dark green), Tenericutes (light blue), Chlorobi (magenta), Fusobacteria (flash green), Thermotogae (brown), Planctomycete (dark orange) and Chloroflexi (orange)*

Differential analysis of whole-genome shotgun sequences*

Sohan Seth [1], Niko Välimäki [2,3], Samuel Kaski [1,3], Antti Honkela [3]
[1]Helsinki Institute for Information Technology HIIT,
Department of Information and Computer Science, Aalto University, Espoo, Finland
[2]Genome-Scale Biology Program and Department of Medical Genetics,
University of Helsinki, Helsinki, Finland
[3]Helsinki Institute for Information Technology HIIT,
Department of Computer Science, University of Helsinki, Helsinki, Finland

1 Introduction

A metagenomic sample comprises genetic material from multiple unknown organisms. In the early years of metagenomic data analysis, researchers were typically interested in learning the composition and relative abundances of these organisms for a single metagenomic sample. Recently, the availability of multiple metagenomic samples from similar environments (e.g., human gut) has allowed the researchers to explore other important issues, e.g., given two metagenomic samples, how similar they are [1], or given two groups of metagenomic samples (cases and controls), what makes them different [2].

In this abstract, we discuss the latter problem, i.e., we study what organisms or sequences are *enriched or differentially expressed* in one group compared to the other. Enrichment analysis or differential analysis of metagenomic samples has recently gained considerable attention [2], and has led to novel understanding of abnormal changes in the natural state of human microbiome under aberrant conditions. For instance, the study of healthy human gut microbiota versus that of patients with obesity, diabetes, or inflammatory diseases has identified many healthy organisms that disappear from the gut in patients afflicted by these diseases, and are often replaced by harmful organisms. This potentially leads to the development of early detection and prevention methods.

Traditionally, the differential analysis of metagenomic samples involved three basic steps: *mapping* the reads to existing markers such as genes or pre-determined unique strings, *estimating* the abundance of the taxa from the mapped reads, and *testing* the enrichment on the estimated abundances using standard statistical tests (see, e.g., [3]). In essence, this approach tackles an arguably more difficult problem of estimating abundance of the organisms where one is only interested in knowing if they are more or less enriched. Additionally, determining the abundance of taxa from pre-determined markers is restricted by the availability of such markers, and is not sensitive to novel variation not present in the reference databases. In order to allow reference-free analysis that is sensitive to novel and small scale genetic variation too, we propose an alternate approach for differential analysis of metagenomic samples. Rather than studying the enrichment at the level of taxa or other predetermined markers, we suggest studying enrichment at the level of short sequence features, k-mers. Briefly speaking, we find k-mers of arbitrary length (upper bounded by

*Part of the calculations presented in this abstract were performed using computer resources within the Aalto University School of Science "Science-IT" project. This work was supported by the Academy of Finland (project numbers 140057, 250345, 251170 and 259440)

read length) that are enriched in one group or the other. Such k-mers form a natural basis for followup analysis by mapping to reference databases or targeted assembly of the reads. We find such k-mers for type II diabetes data and report some preliminary validation.

2 Approach

We summarize our pipeline in Figure 1. Below, we describe the relevant steps in detail.

Distributed string mining: The basic problem of k-mer counting has seen many variations and solutions, most of which solve it for a predetermined (fixed) value of k. We consider a more general approach that allows us to test the abundances over all k-mer lengths (and over all samples) simultaneously. The generalized problem is more computationally demanding but can be solved in feasible time by resorting to a distributed algorithm. We employ the *Distributed String Mining* (DSM) framework [1] and extend it to support our hypothesis testing.

Hypothesis testing: Our goal is to detect k-mers that are differentially expressed in terms of the mean abundance. Let n_{ij}^c be the number of occurrences of a k-mer j in sample i belonging to condition $c = 1, 2$. We model the count distribution of a single k-mer over multiple samples by a negative binomial (NB) distribution to strike a balance between speed and accuracy (other possible distributions can be log-normal, faster to estimate but not discrete, or zero-inflated models, slower to estimate but better motivated). Thus, we utilize the following model for each k-mer count, $n_{ij}^c \sim \mathrm{NB}(n; s_i\mu_j^c, \alpha_j^c)$ where

$$\mathrm{NB}(n; \alpha, \mu) = \frac{\Gamma\left(n + \frac{1}{\alpha}\right)}{\Gamma\left(\frac{1}{\alpha}\right)\Gamma(n+1)} \left(\frac{\alpha\mu}{1+\alpha\mu}\right)^n \left(\frac{1}{1+\alpha\mu}\right)^{\frac{1}{\alpha}}$$

where μ_j^c and α_j^c are the mean abundances and respective dispersion parameters. The null hypothesis is that $\Theta_0 : \mu_j^1 = \mu_j^2$. Here the s_i are universal normalizing constants, same for all j that take into account the different coverage depths of the metagenomic samples, and determined *a priori*. Due to the lack of standard tests (such as the t-test) on negative binomial distribution, we suggest using a likelihood ratio test to determine if one set of observations has a different mean than the other set of observations. Therefore, our test statistic is

$$L = -2\ln\left(\frac{\sup_{(\mu_j^1,\alpha_j^1,\mu_j^2,\alpha_j^2)\in\Theta_0\subset\Theta} \mathrm{NB}(n_{1j}^1,\ldots,n_{C_1j}^1;\mu_j^1,\alpha_j^1)\mathrm{NB}(n_{1j}^2,\ldots,n_{C_2j}^2;\mu_j^2,\alpha_j^2)}{\sup_{(\mu_j^1,\alpha_j^1,\mu_j^2,\alpha_j^2)\in\Theta} \mathrm{NB}(n_{1j}^1,\ldots,n_{C_1j}^1;\mu_j^1,\alpha_j^1)\mathrm{NB}(n_{1j}^2,\ldots,n_{C_2j}^2;\mu_j^2,\alpha_j^2)}\right)$$

where C_1 and C_2 are number of samples in each conditions. However, instead of testing $\Theta_0 : \mu_j^1 = \mu_j^2$ (which involves optimization over three variables μ, α_1, α_2 for the numerator), we suggest testing $\mu_j^1 = \mu_j^2$ and $\alpha_j^1 = \alpha_j^2$ which involves optimization over two parameters μ, α for both numerator and denominator. We use Newton's method to optimize the parameters. If Θ_0 is rejected, we say the k-mer is enriched in condition 1 (2) if $\mu_j^{1(2)} > \mu_j^{2(1)}$.

Reference database: Post-analysis of the enriched k-mers is done by mapping them against the *HMREFG Reference Genome Database*[1]. The mapping is done with exact matches only, our focus being on enriched k-mers that map uniquely to an organism in the database.

[1] http://www.hmpdacc.org/HMREFG/

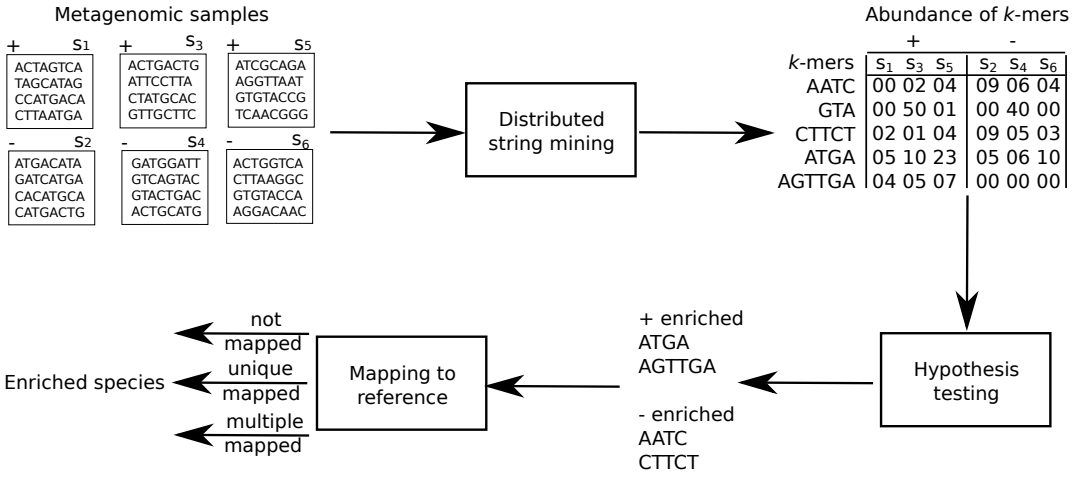


Figure 1: Illustration of the proposed approach which comprises three stages. First, *computing* the abundance of all k-mers efficiently over multiple samples (with *distributed string mining* framework), *testing* the enrichment of each k-mer (with suitably fast *hypothesis testing*), and *analyzing* the enriched (unique) k-mers for finding enrichment at the taxon level (by *mapping to reference database*). We use likelihood ratio test under negative binomial distribution for hypothesis testing. We do not store the intermediate k-mer abundance table, but only the enriched k-mers.

Related methods: Differential analysis is one of the most fundamental techniques in functional genomics. Several tools such as edgeR, DESeq, and metagenomeSeq focus on analysis of count measurements. The former two especially focus on RNA-seq and sharing information across features with very small sample sizes (as small as $n = 3$), while metagenomeSeq focuses on marker gene surveys with a sophisticated zero-inflated model shared across features. They all require access to the full count matrix which is infeasible for us because of the very large number of k-mers we need to process. The scale of the problem restricts us to applying the likelihood ratio test separately on each k-mer in an online fashion *without any knowledge sharing or without explicitly storing the counts as such* which is common to the other approaches. We exploit this situation further and instead of choosing a particular k, we apply the test on *all possible k's*.

3 Results

We investigate our setup on a type II diabetes dataset comprising 199 whole genome shotgun sequencing samples: 99 samples are from diabetic patients, 100 samples are from healthy people [2][2]. Since, enrichment analysis in gut microbiome is an established area of research, this gives us an opportunity to validate our results using existing findings at taxa level. We achieve this by using the fact that sufficiently long k-mers are often unique to a taxon. Therefore, we map the enriched k-mers to a suitable reference database, and investigate the ones that map uniquely. Given a sufficient number of such k-mers for a certain taxon, the enrichment of the taxon can be inferred simply by taking the mean over individual k-mers.

We used a p-value threshold of $\exp(-10)$ to restrict the false discovery rate to (around) 1% (using Benjamini-Hochberg procedure). About 3 million k-mers of varying lengths pass this thresh-

[2]we chose to explore the phase II data instead of the phase I data since the former has higher coverage; about 40% more reads than the latter

ID	Organism	# of unique k-mers total	T2D	CON	k-mer length mean $\pm$ std	log-fold-change mean $\pm$ std
174	Bacteroides coprocola DSM 17136	27870	71	27799	47.35±17.90	-4.18±2.51
372	Clostridium bartlettii DSM 16795	31104	44	31060	44.64±17.09	-3.75±0.87
1001*	Roseburia intestinalis L1 82	2259855	471	2259384	42.62±16.86	-3.34±1.19
546*	Faecalibacterium prausnitzii M21/2	61406	90	61316	31.91±12.85	-2.22±0.78
1002*	Roseburia inulinivorans DSM 16841	18056	259	17797	33.58±16.00	-1.95±1.38
182	Bacteroides pectinophilus ATCC 43243	19427	91	19336	40.36±18.36	-1.91±0.82
545*	Faecalibacterium prausnitzii A2 165	18882	383	18499	29.02±10.67	-1.80±1.09
539*	Eubacterium rectale ATCC 33656	181315	108	181207	38.04±13.96	-1.74±0.87
537*	Eubacterium eligens ATCC 27750	81902	72	81830	46.14±19.74	-1.53±0.46
1023	Ruminococcus obeum ATCC 29174	10667	843	9824	28.12±10.29	-1.33±1.34
1024	Ruminococcus sp. 5-1-39BFAA strain	16475	952	15523	28.39±11.53	-1.25±1.05
396	Clostridium methylpentosum DSM 5476	10828	10485	343	43.07±14.08	2.04±1.14
538	Eubacterium hallii DSM 3353	17282	11784	5498	25.40±8.93	2.63±3.44
1314*	Bacteroides sp. 20-3	23884	23710	174	46.87±18.12	3.05±1.53
374*	Clostridium bolteae ATCC	15877	14631	1246	29.51±13.27	3.80±2.82
171*	Bacteroides caccae ATCC 43185	46884	46389	495	65.27±12.89	3.82±1.73
173*	Bacteroides cellulosilyticus DSM 14838	27089	26705	384	50.01±17.95	4.19±2.02

Table 1: Organisms found enriched in case (patients with diabetes) and control (healthy people) by the suggested method. We observe that the organisms found by us closely match the ones reported in other related publications [2, 3, 4] (denoted by *). It should be noted that the k-mers that map uniquely are sufficiently long (column 6), most of the unique strings for a particular species are enriched in same direction (columns 3,4,5), and the enrichment of individual k-mers for a particular organism agree with each other well (column 7).

old. After mapping these k-mers to the reference database, we consider organisms for which at least 10,000 uniquely mapped strings could be found. We evaluate the log-fold-change as $\log_2((\mu^{\text{case}} + \epsilon)/(\mu^{\text{control}} + \epsilon))$ with $\epsilon = 0.01$ to avoid division by zero and $\log(0)$ error. Thus a positive fold change implies case (diabetes) enriched organism, and a negative fold change implies a control (healthy) enriched organism.

We report the organisms for which the absolute log-fold-change (base 2) is greater than 1 in Table 1. We observe that the enriched organisms found by our setup tally well with other published results. For example, B. cellulosilyticus, B. caccae, C. bolteae, B. sp. 20-3 have been reported to be enriched in diabetic patients [2, 3]. Similarly, R. inulinivorans, R. intestinalis, F. prausnitzii, E. rectale have been reported to be enriched in diabetic patients by [2, 3]. Additionally, E. eligens was reported to be depleted [4]. The full extent and biological significance of the rest of the reported organisms require further study.

References

[1] Sohan Seth, Niko Välimäki, Samuel Kaski, and Antti Honkela. Exploration and retrieval of whole-metagenome sequencing samples. *Bioinformatics*, 2014. doi:10.1093/bioinformatics/btu340.

[2] Junjie Qin, et al. A metagenome-wide association study of gut microbiota in type 2 diabetes. *Nature*, 490(7418):55–60, 2012.

[3] Qichao Tu, Zhili He, and Jizhong Zhou. Strain/species identification in metagenomes using genome-specific markers. *Nucleic Acids Research*, 42(8):e67, 2014.

[4] Fredrik H. Karlsson, et al. Gut metagenome in European women with normal, impaired and diabetic glucose control. *Nature*, 498(7452):99–103, 2013.

Structural inference in oscillatory networks: a case study of the *Arabidopsis thaliana* circadian clock.

Daniel Trejo-Banos[1], Guido Sanguinetti[1,2] and Andrew J. Millar[2,3]
[1]School of Informatics, [2]SynthSys, Synthetic and Systems Biology, Edinburgh,
[3]School of Biological Sciences, University of Edinburgh.

1 Introduction

Oscillations lie at the core of many biological processes, from the cell cycle, to circadian oscillations and developmental processes. Time-keeping mechanisms are essential to enable organisms to adapt to varying conditions in environmental cycles, from day/night to seasonal. Transcriptional regulatory networks are one of the main mechanisms behind these biological oscillations. However, while identifying cyclically expressed genes from time series measurements is relatively easy, determining the structure of the interaction network underpinning the oscillation is a far more challenging problem.

Here, we explicitly leverage the oscillatory nature of the transcriptional signals and present a method for reconstructing network interactions tailored to this special but important class of genetic circuits. Our method is based on projecting the signal onto a set of oscillatory basis functions using a Discrete Fourier Transform (DFT) and using a linear time invariant approximation to the system dynamics. As frequency spectra are invariant under linear transformations, this reduces the problem to a regression problem where the derivative of the signal is regressed against the signal. We build a Bayesian Hierarchical model on top of the frequency domain linear model, in order to enforce sparsity and incorporate prior knowledge about the network structure. Sparsity is induced by selecting a Spike and Slab prior over the network topology. We also briefly explore the integration of additional side information in the form of pairwise promoter sequence similarity scores, as a potential tool to group target genes coregulated by a transcription factor. Our results indicate that our method is significantly better than competitors when the oscillatory assumption is met, and is comparable to alternatives when it is not entirely met.

2 Methods

Our method starts with a linear-time invariant approximation to the system's dynamics (see [1]), resulting in a system of ODEs of the form:

$$\frac{\mathrm{dx_i}}{\mathrm{dt}} = \sum_{\mathrm{j}\neq\mathrm{i}}^{\mathrm{N}} \alpha_{\mathrm{ij}}\mathrm{x_j} + \mathrm{b}_i - \lambda_i \mathrm{x_i} + \sum_k c_{ik}\mathrm{u}_k \tag{1}$$

where the rate of change of gene x_i is given by the sum of activating/repressing actions of N transcription factor's expression levels, the α_{ij} parameters, plus a basal expression level b_i and with a decay rate of λ_i. Additionally system inputs $\mathrm{u_{ik}}$ and input parameters c_{ik} are considered . By

projecting the gene expression data over a set of known basis functions the derivative $\frac{\mathrm{dx_i}}{\mathrm{dt}}$ can be computed analytically. Having X_i being the DFT of the gene expression x_i and its derivative $\dot{\mathrm{X}}_i$, equation (1) in a sinusoidal basis is given by:

$$\dot{\mathrm{X}}_\mathrm{i} = \sum_{\mathrm{j}\neq\mathrm{i}}^{\mathrm{N}} \alpha_{\mathrm{ij}}\mathrm{X}_\mathrm{j} - \lambda_i \mathrm{X}_\mathrm{i} + \sum_k c_{ik}\mathrm{U}_k. \tag{2}$$

Let us define matrix $\mathbf{X}$, each column contains the frequency spectrum coefficients X_i, and let $\mathbf{A}$ be the matrix with diagonal elements λ_i and off-diagonal elements α_{ij}, and $\mathbf{C}$ the matrix of input weights c_{ik}. We now can cast the learning of parameters $\mathbf{A}$ and $\mathbf{C}$ as a regression problem, in which the derivative of the frequency spectrum is regressed against the frequency spectrum of the expression levels and the inputs. The focus will be on inferring the network structure that is embedded in the interaction matrix $\mathbf{A}$. With this purpose, the adjacency matrix $\mathbf{H}$ is defined as the matrix with elements $h_{ij} = 1$ if gene x_i is activated or repressed by gene x_j.

If a set of K time series is available, the likelihood of the parameters for a set of time series spectra$\left\{\mathbf{X}^\mathrm{k}\right\}$, $\mathrm{k} \in [1, \mathrm{K}]$, and their derivatives $\left\{\dot{\mathbf{X}}^\mathrm{k}\right\}$ is

$$\mathrm{P}\left(\left\{\dot{\mathbf{X}}^\mathrm{k}\right\}|\mathbf{A}, \mathbf{C}, \sigma_a\right) \;=\; \prod_{\mathrm{k}=1}^{\mathrm{K}} \mathrm{P}\left(\dot{\mathbf{X}}^\mathrm{k}|\mathbf{A}, \mathbf{C}, \mathbf{X}^\mathrm{k}, \sigma_a\right) \tag{3}$$

Assuming the approximation error to be Gaussian (i.e., the discrepancy between the l.h.s. and the r.h.s. of equation (2) arising from model mismatch and finite dimensional projection), equation (3) will be a product of Gaussian distributions. For simplicity, we will assume that the approximation error will be i.i.d. across different time series, although this assumption can be relaxed trivially.

Given prior assumptions, our goal is inferring the posterior distribution $\mathrm{p}\left(\mathbf{H}|\left\{\mathbf{X}^k\right\}, \mathbf{A}, \mathbf{U}, \mathbf{C}, \sigma_a\right)$. Following a Bayesian approach, we set a prior probability over the matrix $\mathbf{H}$. A desirable characteristic of this prior is to promote sparsity avoiding a fully connected network that would result from estimating $\mathbf{A}$ with ordinary regression techniques. The chosen prior consists of a mixture of a degenerate distribution, the "spike" at zero and a long tailed distribution, the slab. For the proposed model, the specific form of spike and slab prior is a variation of the one presented in [3], and is given in the hierarchical form:

$$\begin{array}{rcc}
\mathrm{P}(a_{ij}|h_{ij},\tau_{ij}) & \sim & h_{ij}\tau_{ij}^{-2}\exp\left(-\frac{a_{ij}^2}{2h_{ij}\tau_{ij}^2}\right) \\
\mathrm{P}(h_{ij}|w) & \sim & (1-w)\delta_{v0}+w\delta_1 \\
\pi(w) & \sim & \mathrm{Beta}(\mathrm{a}_1,\mathrm{a}_2) \\
\pi\left(\tau^{-2}\right) & \sim & \mathrm{Gamma}(\mathrm{b}_1,\mathrm{b}_2) \\
\pi\left(\sigma_a^{-2}\right) & \sim & \mathrm{Gamma}(\mathrm{c}_1,\mathrm{c}_2).
\end{array} \tag{4}$$

Parameters h_{ij} and τ_{ij} jointly define a bi-modal continuous distribution, the "spike" at $v0$ shrinks the values of the a_{ij} coefficients towards it. The choice of parametrization for the Gamma prior over

τ_{ij}^{-2} is such that the resulting distribution has a long continuous tail. The complexity parameter w is drawn from a Beta distribution according to prior knowledge about the network sparsity. Finally a Gamma prior is chosen for the approximation error precision.

Incorporating sequence information. As an additional source of information, we introduce the pairwise similarities between the promoter sequences of the network components. We assume that the overall similarity between two promoter sequences is related to the number of shared regulators. We model the pairwise similarities s_{ij} contained in a similarity matrix $\mathbf{S}$, as normally distributed and proportional to the number of shared regulators, which is obtained from the Hadamard product of rows i and j of $\mathbf{H}$.

The gene expression model and the sequence similarity model are integrated in the same hierarchical model with $\mathbf{H}$ and sparsity parameter w on top of the hierarchy. A Gibbs sampling scheme is used in order to infer the model parameters given the observed gene expression data and pairwise sequence similarities.

3 Experiments and conclusions.

We assess here the performance of our method on two data sets. In both cases, we compare the hierarchical Bayesian method (*sns*) with a Lasso minimisation of the neg-Log Likelihood derived from 3, and with a time series adaptation of GENIE3 [2], a winner method on the DREAM4 competition. We compare performances by plotting Precision Recall (P-R) curves, and their respective area under the P-R curve (AUPR) and Area under the ROC curve (AUC) for comparison.

We first tested our method on a benchmark data set, the DREAM 4 network inference challenge [5]. This competition provided a set of 5 time series for a 10 node network whose dynamics are prescribed by a set of nonlinear ODEs. The expression signals have a damped oscillatory behaviour; thus, there is a considerable level of model mis-specification in this case. Figure(1) top right panel shows the respective P-R curves (including a flat random baseline). Despite the model mismatch, we see that sns improves over both GENIE3 and the Lasso solution.

We then tested the method on an oscillatory model for the *Arabidopsis thaliana* Circadian clock Network [4]. This nonlinear model contains 6 known transcriptional components of the *Arabidopsis* circadian clock network and a hypothetical regulator Y, excluding modified proteins LHYmod and TOC1mod (Fig. 1 bottom left). The model generates regular oscillations driven by a light input. Simulations corresponding to 4 different photo periods, and 4 different mutant types were produced. All these time series were down sampled in order to resemble experimental conditions. The true adjacency matrix was derived from the model formulation and used as ground truth. We also used a small linear model for simulating similarity matrix $\mathbf{S}$ and tested adding the simulated sequence information to the inference. Results are presented in figure(1).

In this case, the sns method showed a considerable improvement in PR space, both relative to Genie3 and the lasso solution, with an AUPR of 0.65. By adding sequence information, performance is excellent, with an AUPR of 0.88.

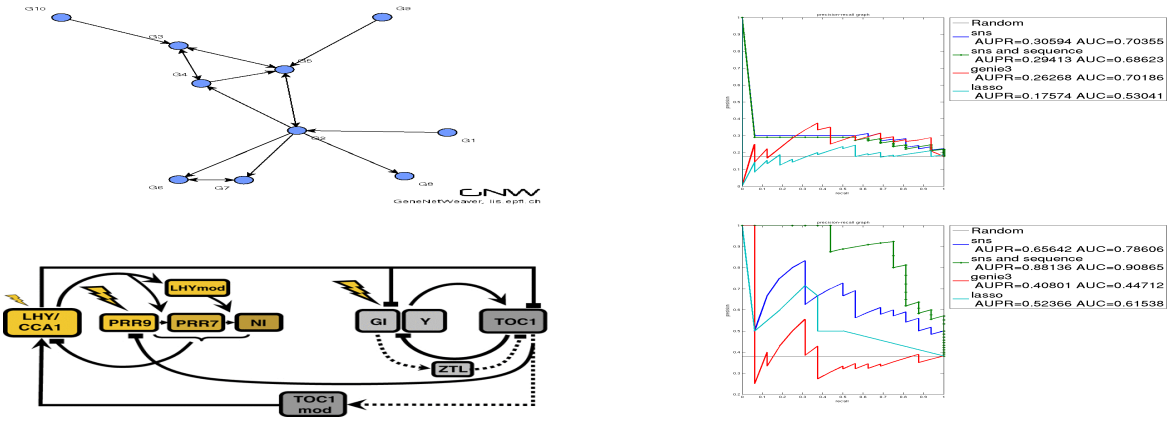


Figure 1: Top left, Dream4 oscillatory network, bottom left Circadian clock model[4]. Top right and bottom right, the corresponding PR curves for the sns method, the sns with promoter sequence addition, the genie3 method and the lasso solution.

These preliminary experiments demonstrate that our method can indeed give considerable advantages over competitors when the system under study is indeed oscillatory, and is relatively robust to model mismatch. Our method shares the computational limitations of other Bayesian model-based approaches: in the experiments we showed, 5000 Gibbs samples were used to estimate posterior distributions, with a total computational cost of 150 seconds. Our method also retains the advantages of Bayesian methods, given by a more transparent interpretation and a principled quantification of uncertainty; depending on the situation, the high computational costs may well be a price worth paying.

References

[1] N. Dalchau. Understanding biological timing using mechanistic and black-box models. *New Phytologist*, 2012.

[2] V.A. Huynh-Thu et al. Inferring regulatory networks from expression data using tree-based methods. *PLoS ONE*, 5(9):e12776, September 2010.

[3] Hemant Ishwaran and J. Sunil Rao. Spike and slab variable selection: Frequentist and bayesian strategies. *The Annals of Statistics*, 33(2):730–773, April 2005.

[4] A Pokhilko, et al. Data assimilation constrains new connections and components in a complex, eukaryotic circadian clock model. *Molecular Systems Biology*, 6, September 2010.

[5] T. Schaffter et al. GeneNetWeaver: in silico benchmark generation and performance profiling of network inference methods. *Bioinformatics*, 27(16):2263–2270, August 2011.

Interspecies Association Mapping: connecting phenotypes to sequence regions across species

A.D.J. van Dijk, Bioscience & Biometris, Wageningen University and Research Centre, The Netherlands; aaltjan.vandijk@wur.nl

Introduction

Genome Wide Association Studies (GWAS) capitalize on large amounts of available trait and sequence data for individuals within a species, and indicate specific sites of sequence variation that are associated with trait variation. A basic assumption in a conventional GWAS is that the genome sequences of individuals involved in the analysis can be directly compared. Any differences between genomes of different individuals, such as Single Nucleotide Polymorphisms, can then be used as explanatory variables that potentially influence the phenotype.

Zooming out from the level of any particular species, for many different species sequence data are available, and the amount of that data grows rapidly. Also, a lot of phenotypic data is available describing many different species. It would be of great biological interest to find specific sequence sites that influence differences between species. However, there is no method to analyse sequence and trait data between species, and find sequence sites involved in trait differences. A method simply analogous to the GWAS type of approach is not applicable, because of the very different sequence content of different species; it is not possibly to use the differences between genomes of different species as predictors.

Here, we provide an approach towards 'interspecies association mapping' (Fig. 1). Briefly, it first performs sequence clustering based on PFAM domain composition. For each PFAM domain architecture, a separate classification model is trained, using the trait value as dependent variable. Based on performance in predicting the trait value, relevant PFAM domains are selected. Subsequently, specific relevant sites in sequences with such PFAM domains are selected based on the importance of the features used for classification.

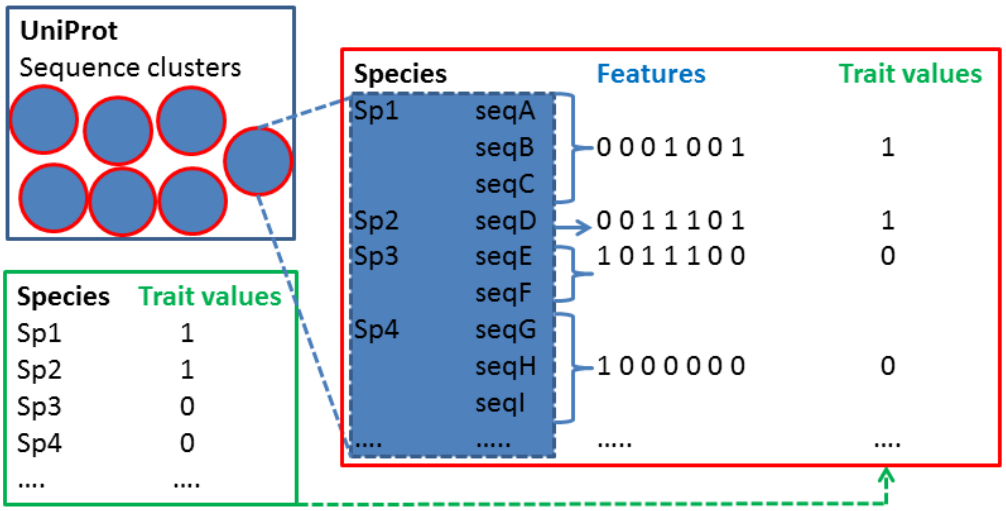


Figure 1. Overview of method for interspecies association mapping. Trait values for many different species (green box) are combined with sequence data from UniProt (blue box). The latter are clustered based on PFAM domain architecture. For each cluster of sequences with a particular PFAM domain architecture, a classification model is trained (red box) using features per species which are defined based on the set of sequences in the sequence cluster that belong to that species. For those classification models (PFAM domain architectures) which result in classification performance better than based simply on sequence similarity, relevant features are selected based on the importance of the features in classification. Those selected features indicate specific sequence positions relevant for variation of traits between species.

Method

Four different plant traits were used, each with discrete values (Table I). All protein sequences from UniProt for plant species were obtained, and grouped based on PFAM domain info provided by UniProt. These sequences were used to define features, and train a support vector machine (SVM) classifier, as described below. For the PFAM domains that were best performing in the train/test setup based on UniProt data, additional sequences were obtained via NCBI blast. These were used for further validation: again features were defined and the model trained on the UniProt data was applied.

Table I. Trait data used in interspecies association mapping[a]

Trait	Trait values	#species	% minor class	datasource
Bloom period	Spring vs Summer	1850	Summer: 34%	[1]
Duration	Annual vs Perennial	31398	Annual: 20%	[1]
Leaf phenology	Deciduous vs evergreen	17266	Deciduous: 17%	[2]
Leaf compoundness	Simple vs compound	35085	Compound: 14%	[2]

[a]Four different discrete traits were used, for which trait values were available for the number of species indicated. The percentage of species for which the trait has the minor class value is indicated as well.

We took as a starting point the conjoint triad string kernel, as previously proposed in the context of protein interaction prediction [3]. This involves a set of 8000 (20x20x20, where 20 is the number of amino acids) different triads. In a general setting, feature *n* for protein sequence S would have value 1 if and only if the triad represented by that feature is a substring of the protein sequence, otherwise it would have value 0. However, a particular aspect of the case at hand is that for each unit in the classification problem (a species with its trait-value) there are one or more sequences associated to the species, which are used as input for defining the features. Hence, in this case, feature *n* for species S has value 1 if and only if the triad represented by that feature is a substring of at least one of the protein sequences present for species S; otherwise it has value 0.

As SVM implementation SVMlight [4] was applied with a radial basis function (RBF) kernel. Cross-validation was applied to optimize the parameters C (trade-off between training error and margin) and γ (RBF parameter), for which a grid was used (values of [30 20 15 10 5 2 1] and [500 200 100 10 5 1 0.1 0.01 0.001 0.0001], respectively). To obtain an unbiased performance estimate, a nested cross-validation setup, as described previously, was used [5]. Note that this setup avoids erroneously optimistic estimates obtained by simply using cross-validation to optimize the SVM parameters. Within the inner cross-validation loop, best performing parameters were selected using F-score, the harmonic mean of precision (=TP/(TP+FP)) and recall (=TP/(TP+FN)). Here, TP is true positive, FP is false positive, and FN is false negative; positives are defined on the smallest class. F-score was expressed on a scale from 0 to 100%. Selected parameter values were applied on the test set (outer cross-validation loop) to get an unbiased performance estimate. To analyse feature importance, a Sensitivity Index was used as described previously [6].

As benchmark to compare the SVM performance with, sequence similarity based predictions were performed. For the sequence similarity based predictions, only the longest sequence per species was used in order to be able to align the large number of sequences. Alignment was performed using MUSCLE [7]. For each species, a trait prediction was obtained using this alignment by identifying the species with the sequence with the highest identity and transferring the class label of that species.

Results

We built classification models to predict properties of species, using four different plant traits: bloom period, which distinguishes plants that flower early (in spring) or late (in summer); duration, which distinguishes annual plants (flowering once and completing their lifecycle in just one season) vs perennial plants (flowering several years); leaf phenology, which describes whether a plant loses its leaves for part of the year (deciduous) or not (evergreen); and leaf compoundness, which distinguishes plants with simple leaves (single blade in a leaf) vs those with compound leaves (the blades of which are divided into distinct parts).

Results for predicting those species properties based on sequence features were assessed using an F-score, compared to the F-score of simple sequence similarity based prediction ("background model"). Predictions were performed separately for different PFAM domain groups. For both bloom period and duration, there were two well performing PFAM cases separated from the background (Table II, Figure 2): ATP synthase domains and TCP transcription factor domain for bloom period, and Hsp20/alpha crystallin family and the C-terminal Glyceraldehyde 3-phosphate dehydrogenase domain for duration. For leaf phenology and leaf compoundness, no PFAM domain was selected with a clear distinction in performance against the background (Table II, Figure 2).

For bloom period, the role of ATP synthase is unclear. However, the second selected domain, TCP, is quite relevant indeed. For example, TCPs influences flower maturation [8, 9], and promote CK [10], which influences flowering time as well [11]. For duration, it is unclear whether the Hsp domain is relevant, although it might be that heat-shock (in which the Hsp domain is involved) is indeed relevant [12]. For Glyceraldehyde 3-phosphate dehydrogenase, there is a clear link to the trait via the importance of carbohydrate reserves [13].

Table II. Results of PFAM domain selection based on performance in classification

Trait	#PFAM input sets[a]	Average #species[b]	Selected PFAM[c]
Bloom period	69	97+/-103	PF03634 (TCP)
			PF00006/ PF02874 (ATP synthase)
Duration	205	204+/-374	PF00011 (Hsp20/alpha crystallin)
			PF02800 (Glyceraldehyde 3-P dehydrogenase)
Leaf phenology	110	188+/-395	NA
Leaf compoundness	235	227+/-541	NA

[a]Number of PFAM-based clusters, for each of which a predictive model was built. Cutoff on minimum number of different species was applied (>50 species).
[b]Average (+/- stdev) number of species in the PFAM-based clusters for which a model was built.
[c]PFAM domain clusters for which the SVM model obtained a performance better than the background model (based on sequence similarity).

Importantly, the above mentioned performance estimates were obtained using completely unseen test-sets. Nevertheless, for further validation, additional sequences containing the selected PFAM domains were obtained for both bloom period and duration. Although the additional number of sequences was limited, the validation data showed reassuring performance (Table III).

Table III. Performance in cross-validation (test-set) and in validation set[a]

Trait	Selected PFAM	Test-set F-score	#new species in validation set	F-score validation set
Bloom period	PF03634	0.58+/-0.40	6	0.67
	PF00006/PF02874	0.31+/-0.39	5	0.50
Duration	PF00011	0.73+/-0.18	16	0.70
	PF02800	0.71+/-0.16	97	0.55

[a]For the traits for which classification models were obtained with clearly better performance than sequence based predictions, data from additional species were obtained and the performance (F-score) of predicting the class for those additional validation data was assessed.

Our approach enables to find specific sequence regions relevant for trait variation between species. We validated those specifically for the TCP domain, which was selected for Bloom period. Here, in the TCP domain there was an overlap of the most important features with known DNA binding residues [14]. For example, in the maize protein B6SSJ3_MAIZE, the first part of the TCP domain consists of GGHIVRSTGRKDRHSKVCTARGP**RDRR**V**RLS**; here the higlighted residues indicate the residues selected by our method. In this region, the change of RDR to RGR determines preference for specific DNA binding site, which correlates with evolutionary properties in eurosids [15]. Also, in Arabidopsis TCP11 a Thr instead of Arg in the RLS site influences strongly the DNA binding preferences [16].

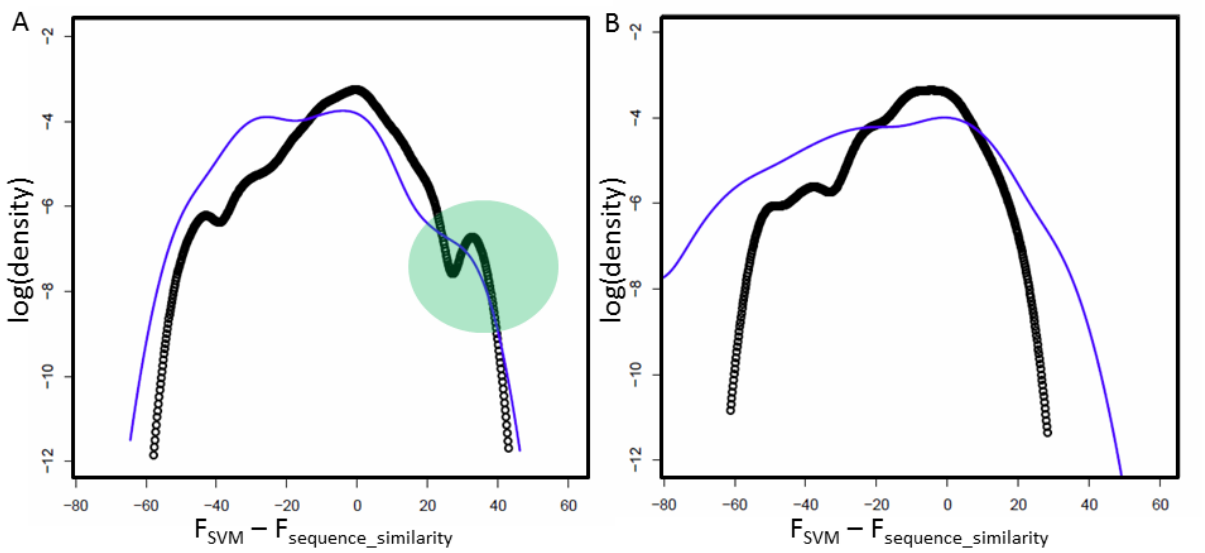


Figure 2. Classification performance. Log-density of SVM-based F-score (F_{SVM}) minus sequence similarity based F-score ($F_{sequence\_similarity}$) for the different PFAM input datasets. Positive values mean that the feature-based classification performs better than overall sequence similarity based classification. **(A)** Bloom period (black) and duration (blue). **(B)** Leaf phenology (black) and leaf compoundness (blue). Note that for bloom period and duration, there is a 'shoulder' at the right hand side of the distribution (green shading); this indicates the PFAM domain architectures for which the SVM based prediction has a clearly better performance than the background.

Discussion and conclusion

One aspect to investigate further is how strongly trait values are correlated with phylogeny. We gauged the predictive performance of our SVM model as compared to overall sequence similarity based models in order to take this into account. For one trait, leaf compoundness, it seems that the trait values within a genus are very similar, whereas between closely related genus there is not much similarity. For this particular trait, clustering input data per genus before classification could be advantageous.

Our approach will not be able to pinpoint cases of gene regulatory evolution. However, for two out of the four different traits we tested, we obtained evidence of the importance of coding evolution for trait evolution. The next step would be to apply our approach to additional trait and sequence datasets, including those that are available outside the plant field.

References

1. plants.usda.gov.
2. http://try-db.org/TryWeb/Home.php.
3. Shen, J.W., et al., *PNAS*, 2007. **104**(11): p. 4337-4341.
4. Joachims, T., in *Advances in Kernel Methods - Support Vector Learning*. 1999, MIT-Press.
5. Varma, S. and R. Simon, Bmc Bioinformatics, 2006. **7**: p. -.
6. Zavaljevski, N., F.J. Stevens, and J. Reifman, Bioinformatics, 2002. **18**(5): p. 689-696.
7. Edgar, R.C., Nucleic Acids Research, 2004. **32**(5): p. 1792-1797.
8. Sarvepalli, K. and U. Nath, Plant J, 2011. **67**(4): p. 595-607.
9. Rubio-Somoza, I. and D. Weigel, PLoS Genet, 2013. **9**(3): p. e1003374.
10. Steiner, E., et al., Plant Cell, 2012. **24**(1): p. 96-108.
11. D'Aloia, M., et al., Plant J, 2011. **65**(6): p. 972-9.
12. Xu, Y., C. Zhan, and B. Huang, Int J Proteomics, 2011. **2011**: p. 529648.
13. Bertrand, A., et al., J Exp Bot, 2003. **54**(388): p. 1721-30.
14. Aggarwal, P., et al., Plant Cell, 2010. **22**(4): p. 1174-89.
15. Viola, I.L., et al., J Biol Chem, 2012. **287**(1): p. 347-56.
16. Viola, I.L., et al., Biochemical Journal, 2011. **435**: p. 143-155.

Application of Geometric Kernel Data Fusion in Protein Fold Recognition and Protein Sub-nuclear Localization

Pooya Zakeri[1], Ben Jeuris[3] , Raf Vandebril[3] , Yves Moreau[1],[2]

[1]Department of Electrical Engineering (ESAT), STADIUS Center for Dynamical Systems, Signal Processing and Data Analytics, KU Leuven, [2]iMinds Medical IT and [3]Department of Computer Science, KU Leuven, 3001 Leuven, Belgium

Abstract: Finding an efficient technique for integrating heterogeneous biological data sources has received growing attention. In particular, kernel methods are an interesting class of techniques for data fusion. We look into the possibility of using the geometric mean of matrices instead of the arithmetic mean for multiple kernel learning (MKL). While computing geometric means of matrices is challenging, it hints at an intriguing research direction in data fusion. Geometric kernel fusion (GKF) is used for protein fold recognition and sub-nuclear localizations. It offers a significant improvement over MKL approaches. The experimental results demonstrate that GKF can effectively improve the accuracy of the state-of-the-art kernel fusion model for protein fold recognition and predicting protein sub-nuclear prediction.

Moreover, the limitation of convex linear combinations in coping with integration of different protein features that carry complementary information is investigated. Our proposed fusion frameworks, by contrast, can be used to detect these features with complementary information, which provides an insightful contrast for combining different features of other problems in bioinformatics.

***Keywords*:** Genomic Data Fusion, Protein Fold Recognition, Predicting Protein Sub-nuclear localization, Geometric Kernel Fusion, Geometric Mean, Multiple Kernel Learning

GENERAL GUIDELINES

Various sequence-based protein features and often machine learning methods have been developed to solve Bioinformatics tasks. More attention needs to be paid to finding an efficient and cost-effective method for fusing these different discriminatory data sources for these problems.

This study presents kernel-based computational frameworks for fusing heterogeneous biological data sources by taking more involved geometry means of their corresponding kernel matrices instead of convex linear combinations. It has been observed that geometric data fusion is less sensitive in dealing with complementary and noisy kernel matrices compared to typical multiple kernel learning approaches [1]. Since biological kernels often encode the complementary characteristics of biological data, this motivates us to see the application of geometric data fusion in bioinformatics tasks.

We address two challenging problems in bioinformatics: (1) Protein fold recognition, and (2) Predicting protein sub-nuclear localization. Both problems are among the most essential objectives in cell biology, molecular biology and proteomics. In particular, protein fold and protein sub-nuclear information can help to understand better cellular process and molecular function in a cell.

We consider various sequence-based protein features including pseudo amino acid composition, predicted secondary structure, as well as information extracted directly from position-specific scoring matrices and local alignment kernels. GKF frameworks are employed to integrate these complementary biological data sources. Moreover, our computational model has been developed by incorporating the functional domain composition of proteins through a kernel-based hybridization model.

Our GKF frameworks offer a significant improvement for both protein fold recognition and protein sub-nuclear localization prediction. It is observed that using GKF, we achieve an overall protein fold accuracy of 86.86% on SCOP PDB-40D benchmark dataset [2] and 96.86 on newDD benchmark dataset [3], which represent 20.8% and 7.6% improvement over the state-of-the-art protein fold predictor respectively. Experimental results on the SNL9 benchmark dataset [4] also show an improvement of 4.1% over the state of the art in protein sub-nuclear localization.

INTRODUCTION

Early and late integration are typical approaches to integrate various protein data sources. In addition, the heterogeneous biological data sources can be integrated intelligently and efficiently using partial integration, such as kernel-based data fusion approaches. Using kernel methods is an interesting strategy because it decouples the original data from the machine-learning algorithms by using a representation of the data as a kernel matrix. Symmetric positive definite (SPD) kernel matrices are in fact the nonlinear extension of covariance/correlation matrices and encode the similarity between examples in their respective input space. This demonstrates that the heterogeneous data can all be replaced by appropriately scaled kernel matrices with the same size, and consequently the data heterogeneity vanishes. Then other classification, clustering and prioritization algorithms can access the same data, which is currently not

possible. Indeed, constructing the same representation for all data sets and integrating these representations systematically is the main insight behind kernel fusion methods.
There are several methods for obtaining a valid and fitting kernel by tuning the kernel matrices weights [5]. Finding such weights from training data and replacing the single kernel by a convex linear combination of weighted base kernels is usually referred to as MKL. These weights reflect the relative potential importance of the different data sources in the combined kernel. Several MKL approaches have been proposed in the literature [6-9], during the last decade. Several complicated convex optimization-based approaches have been proposed [6-9], to improve the efficiency of kernel data fusion. They try to optimize the kernel weights based on different optimization criteria. However, on the one hand, the optimized weights of base kernels convex combination of kernel matrices often leads to mixed results and causes an improvement in performance only when dealing with redundant or noisy kernel matrices [9]. Indeed, this type of averaging has usually sensitive behavior in coping with kernel matrices containing complementary and non-redundant information and fails to completely capture all the information for these kernels. Since genomic kernels often encode the complementary characteristics of biological data, applying convex linear combination of base kernels is not very appropriate for biological application.
On the other hand, it has been showed that even the results obtained by employing uniformly weighted kernel fusion are comparable to the results of the best existing MKL approaches in general applications. This is also supported by the equal weights theorem [10], which states when all optimized weights are uniformly distributed on the interval [0.25; 0.75], the performance is hardly changed using equal weights.
Using the Euclidean distance on a convex cone whose interior contains all SPD matrices $P(n)$, we can obtain the arithmetic mean which is uniformly weighted average of the base kernels. By contrast, since it has been illustrated that this type of averaging fails to completely capture all the information for kernels containing complementary, non-redundant information, Euclidean distance on SPD matrices might not be appropriate. Moreover, SPD matrices form a convex cone and not a vector space. This has an effect on the "natural" geometry of SPD matrices, which may not be Euclidean, but rather should rely on concepts from Riemannian geometry. This inspires us to consider other means between SPD matrices that are not relative to the Euclidean distance on $P(n)$ and necessarily a linear of SPD matrices [1]. For example, the mean corresponding to Riemannian distance on $P(n)$ is the geometric mean.

METHODS

GEOMETRIC KERNEL FUSION

In this study, we design, and develop several methods to integrate kernel matrices by taking more involved, geometry inspired means of these matrices instead of convex linear combinations [1]. Such averaging of the base kernels can be interpreted as a kind of integration that expresses the nonlinear relationship between the individual kernels. In particular, we focus on taking the matrix geometric mean of base kernels. For a general number of matrices, the fused kernel is obtained by taking the geometric mean

$$K = F(K_1, K_2, \ldots, Kn) = G(K_1, K_2, \ldots, Kn)$$

Similar to the arithmetic mean (AM), other types of means of SPD matrices such as the harmonic mean (HM), Log-Euclidean mean (LogEM) [11], and geometric mean(GM) result in SPD kernels. However, computing the geometric mean of a general number of SPD matrices is a challenge. In fact, for a general number of SPD matrices, a proper definition of a geometric mean with some natural properties has only recently been developed [12]. We present two methods for computing the geometric mean [1]. The first approach is focused on computing the actual geometric mean using the definition of the Karcher mean (Karcher-KF) [13]. The second, however, only computes a rough approximation of the actual geometric mean using a proposed, heuristic method based on Arithmetic-Geometric-Harmonic (AGH) mean (AFH-KF). We show in the second section that it is a computationally scalable method for computing an approximate geometric mean.
However, computing the geometric mean for a general number of SDP matrices is hard and computationally expensive, which is why we also discuss the Log-Euclidean mean (LogE) [11] . It can be considered as a consensus between the arithmetic and geometric mean. The Log-Euclidean mean of n SPD matrices can be obtained explicitly as [11]:

$$K_{LE}(K_1, \ldots, K_n) = \exp\left(\frac{1}{n}\left(\sum_{i=1}^{n} \log(K_i)\right)\right)$$

Protein Fold Recognition and Predicting Protein Sub-nuclear Localization

Protein fold recognition is a challenging problem in computational biology. Tertiary structural information of proteins can provide new knowledge on their function. In addition, understanding the three-dimensional structure of proteins can be facilitated through the knowledge of protein folds; hence, determining this structure is among the most essential

objectives in molecular biology, cell biology, and proteomics. Structural information is also potentially useful for drug design study.

Predicting protein sub-nuclear localization is another challenging problem in bioinformatics. Knowledge on functions of proteins can be provided by information of subcellular locations. Another area which has received little attention in the literature [4, 14, 15], is the prediction of protein localization in the organelles of the cell, such as nucleus, chloroplast and mitochondria. The nucleus is a membrane-enclosed organelle in eukaryotic cells, and contains the DNA organized into chromosome. It is the principal location of DNA and RNA synthesis. Information about the sub-nuclear location for a nuclear protein can also provide much better understanding about its function. Furthermore, using the reliable automatic sub-nuclear localizer the design of appropriate drugs could be sped up for many kind of complex diseases linked to human genome and cancers.

In the recent years, various sequence-based protein features and often machine learning methods have been proposed for protein fold recognition [1, 2, 3, 17], and predicting protein sub-nuclear localizations [4, 14]. In order to evaluate the efficiency of the GKF in bioinformatics tasks, we address these two problems.

Feature Vectors

For protein fold recognition, we consider 26 sequence-based protein features including 6 types of structural information (Amino Acid composition (C), Predicted Secondary Structure (S), and 4 pseudo amino acid composition (PseAAC) [4] with four different *landa*, 4 kinds of physicochemical properties of constituent amino acids (Hydrophobicity (H), Polarity (P), van der Waals volume (V) and Polarizability), and 2 local pairwise sequence alignment-based feature spaces, as well as sequence evolution information extracted directly from position specific scoring matrices in 14 different ways.

For predicting protein sub-nuclear localization, we also consider various sequence-based protein features including 2 types of information extracted directly from position specific scoring matrices, 3 types of structural information and local sequence alignment.

RESULTS & DISCUSSION

Gaussian RBF kernel function is employed for different protein features. To see the advantage of fusing heterogeneous data sources for protein fold classification (protein sub-nuclear localization prediction) through intermediate-based data integration, we focus on combining 26 (10) RBF kernel matrices derived from each view on protein domains (nuclear proteins).

The kernel matrices are combined through various types of means like Karcher-KF, AGH-KF, AM, and LogE. Subsequently, the combined kernel is used to determine the performance. Then the classification is performed using a one-versus-rest (OVR) support vector machine (SVM).

We evaluate our method for classification on the SCOP PDB-40D benchmark dataset [2] which consists of 27 SCOP fold classes. Moreover, to compare the performance of our proposed approaches, we also consider three types of MKL approaches [8, 16, 18]. Figure 1 provides the success rates of our proposed kernel fusion approaches based on averaging of the kernel matrices, these MKL approaches, as well as the best existing methods for classification of protein folds in DD data set [1, 2, 3, 16, 17] .

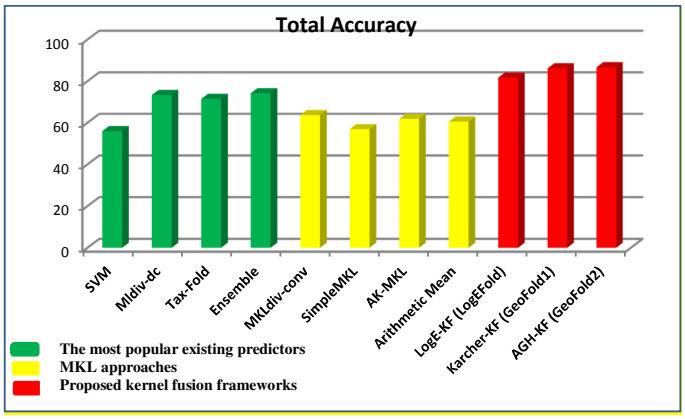


Figure 1. Comparison of the total accuracies of our proposed kernel fusion frameworks with existing predictor and Meta-predictors for classification of protein folds in the DD dataset[2] .

According to Figure 1, classification results of the combined kernels using Karcher-KF, AGH- KF, and LogE-KF show a clear improvement compared to the state of the art.

Furthermore, to incorporate the available functional domain information (FunD) of proteins, we consider the FunD composition of protein sequences using integrated FunD databases. For this purpose, we use the InterPro database [19], and Conserved Domain Database (CDD) [20]. Next, our computational model has been developed by incorporating the functional domain composition of proteins through a hybridization model 9 (Figure 1). It is observed that by using our proposed hybridization model the protein fold recognition accuracy is further improved to 89.30%.

The performance is estimated on an independent test set. The general architecture of the proposed approaches for classifying protein folds and predicting protein sub-nuclear localization are shown in Figure 1.
Furthermore, we investigate the performance of our approach on the newer SCOP database (version 1.75) [3] using 10-fold cross validation. Our results suggest that only by combining the evolutionary and secondary structural kernels through GKF, we obtain competitive results compared with the best ensemble approaches proposed for this problem [17]. In addition, it is observed that by incorporating the available functional domain information through our proposed hybridization model, we are able to tackle the protein fold recognition problem for 27-folds.

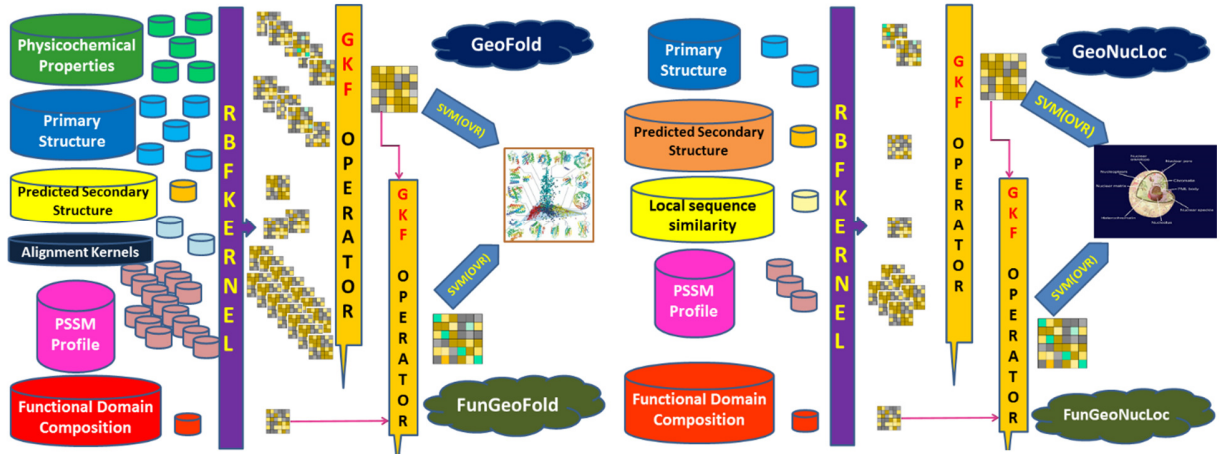


FIGURE 2. (a) The architecture of our fusion model for protein fold recognition (b) and predicting protein sub-nuclear localization.

We also investigate the performance of our geometric kernel fusion approach for predicting protein sub-nuclear localization. We evaluate our models by applying them to the SNL9 dataset [4] data sets using leave-one-out cross-validation. This data set contains 714 nuclear protein classified into 9 sub-nuclear locations. We achieve competitive results compared with the best ensemble approaches proposed for predicting protein sub-nuclear localization.
Moreover, our computational model has been developed by incorporating the functional domain information through the hybridization model. Experimental results on the SNL9 benchmark data set [4] demonstrate that our kernel-based integration approach can effectively improve the accuracy of the state-of-the-art protein sub-nuclear localization predictor.

CONCLUSIONS

In this study, our experimental results demonstrate that the geometric mean of base kernels can effectively improve the accuracy of the state-of-the-art kernel fusion model for protein fold recognition. Moreover, we enhance protein sub-nuclear localization prediction results on the SNL9 data set through our proposed kernel data fusion framework based on the geometric mean of kernel matrices. In addition, incorporating the available knowledge on functions of protein domains into our kernel data fusion framework offers a clear improvement in empirical performance.
Furthermore, the limitation of convex linear combinations in coping with combining different protein features which carry complementary information is considered [1]. By contrast, our proposed fusion frameworks can be used to detect these features with complementary information, which provides an intuitive strategy for integrating different features of other problems in bioinformatics. In particular, our results suggest that integrating the evolutionary and secondary structure information could be crucial to understand the relationship between primary and tertiary structure in proteins.

REFERENCES

1.Zakeri, P., *et al*., *J. THEOR. BIOL*, 269(1), 208 – 216 (2011).
2.Ding, C. H. and Dubchak I., *Bioinformatics,* 17(4), 349-358 (2001).
3.Yang, J.-Y. and Chen, X. *PROTEINS,* 79(7), 2053–2064. (2011)..
4.Shen HB, and Chou KC, *Protein Eng Des Sel*, 20: 561–56712, 2211–2268. (2007).
5.G¨onen, M. and Alpaydin, E. . *J. Mach. Learn. Res*, 12, 2211–2268. (2011).
6.Lanckriet, G. R. G., *et al*., *J. Mach. Learn. Res*, 5, 27–72. (2004).
7.Sonnenburg, S., *et al., J. Mach. Learn. Res*, 328-347 (2006).
8.Rakotomamonjy A., *J. Mach. Learn. Res,* **9**, 2491–2521, (2008).
9.Lanckriet, G. R. G., *et al. Bioinformatics,* 20(16), 2626–2635.(2004).
10.Wainer, H. *Psychological Bulletin*, 83, 213–217. (1976).
11.Arsigny, V. *et al., SIAM J Matrix Anal A,*29(1), 328-347 (2007
12.Bhatia, R., "*Positive Definite Matrices*", Princeton University Press, Princeton Series in Applied Mathematics. (2007).
13.Jeuris, B., *et al., ELECTRON T NUMER ANA,* 39, 379–402 (2012).
14.Lei ZD, and Dai Y., *BMC Bioinformatics*, 6: 291 (2005)
15.Zakeri, P., *et al*., *J. THEOR. BIOL*, 269(1), 208 – 216 (2011).
16.Ying, Y., *et al. BMC Bioinformatics*, 10(1), 267. (2009).
17.Lin, C., *et al, PLoS ONE*, 8(2), e56499. (2013).
18.Qiu, S. and Lane, T. *IEEE/ACM Trans. Comput. Biol. Bioinformatics*, 6(2), 190–199. (2009).
19.Hunter S *et al. Nucleic Acids Res,* 40(D1), 306-312 (2012).
20.Marchler-Bauer, A., *et al*., *Nucleic Acids Research*, 41(D1), D348–D352.. (2013).

Poster abstracts

List of posters

Transcription factor binding site characterization by Bayesian network model averaging

Ankit Agrawal, Rajarshi Pal, and Rahul Siddharthan

The Institute of Mathematical Sciences, Chennai, India

We describe a method to describe transcription factor binding sites by Bayesian networks that specify intra-site dependencies as conditional probabilities. Using Markov chain Monte-Carlo sampling, a set of putative networks, each with Bayesian posteriors given the training data, is obtained, and these are used to calculate the likelihood of a new sequence using model averaging. Ways to visualise the dependencies, similar to "sequence logos" in position weight matrices, are discussed, and performance on synthetic and real biological data is described.

Modeling cancer progression based on maximum a posteriori inference of directed acyclic graphs

Jonas Behr and Niko Beerenwinkel

Department of Biosystems, ETH Zürich, Switzerland

The progression of cancer may be viewed as an evolutionary process in which certain capabilities need to be achieved by the cancer tissue to evade the host immune system and continue growth. It is widely excepted that there is a dependency structure between genomic or epigenomic events, which result in aforementioned capabilities. However, the inference of this dependency structure is challenging and previous approaches ether limited the class of structures or had severe limitations on the number of events they could consider. We have developed a statistical model for the dependency structure of cancer progression based on directed acyclic graphs (DAGs) similar to previous approaches called conjunctive bayesian networks (CBNs). However, certain model simplifications allow us to accurately infer the maximum a posteriori solution using mixed integer programming. This new approach is now able to robustly infer DAGs on more than 200 events, which corresponds to a more than ten fold upscale as compared to previous CBN implementations. At this scale we are now ready to apply the approach to the large scale cancer genomics datasets like TCGA taking all frequently mutated cancer genes into account. We will show a comparison of the cancer progression structures for 20 common cancer types at the meeting.

Predicting binding affinities between drug compounds and kinase targets

Anna Cichonska[1,2*], Tapio Pahikkala[3], Antti Airola[3], Juho Rousu[2], Tero Aittokallio[1]

[1]Institute for Molecular Medicine Finland FIMM, University of Helsinki, Finland

[2] Helsinki Institute for Information Technology HIIT,
Department of Information and Computer Science, Aalto University, Finland

[3]Department of Information Technology, University of Turku, Finland

Introduction Protein kinases are enzymes that play an important role in the cellular regulation by transferring phosphate groups from high-energy donor molecules to particular amino acids of substrate proteins. Deregulated kinase activity is a common cause of diseases. It is well known that these enzymes constitute key regulators of cancer survival pathways. Therefore, effective kinase inhibitors are being designed. Such drugs are small chemical compounds that work by binding to specific kinase targets and blocking their activity. However, determining interactions between drugs and their molecular targets experimentally is time consuming and expensive. In the recent years, a lot of effort has been placed on developing in-silico methods facilitating the process of drug discovery. Fast and efficient machine learning approaches are preferred in the early stage of computational drug screening.

Aim of this work was to predict unknown, not experimentally measured interaction affinities of drug-target pairs based on the large-scale biochemical assay of kinase inhibitors selectivities (Metz et al. [1]). We focused on the regression problem, where the objective is to predict the quantitative binding affinities, instead of more common binary setting, since molecular interactions are not simple on-off relationships [2].

Materials and Methods Metz et al. data set consists of 201 compounds and 169 kinases. On average, 47% of the interactions are missing for a drug. In order to predict unknown drug-target interactions, we used machine learning algorithm utilising Kronecker kernel for regularized least-squares regression (Kronecker RLS) [3]. The basic assumption is that similar compounds are likely to interact with similar targets. Similarities between drugs and similarities between targets, computed using available information, can be encoded using kernel functions and used as features. Kronecker RLS employes a product of drug and kinase kernels. Matrices are combined into a larger kernel that directly relates drug-target pairs. It helps to find important co-occurring features predicting the interactions. We used different ways of computing molecules' similarities. For targets, we used two- and three-dimensional structures' alignments. In case of 2D structures, we computed a kernel also based on extended targets' profile. We applied a linear kernel on the features derived from calculating similarities of given kinases with the bigger set of proteins. In case of drugs, we computed Tanimoto kernels based on 12 different types of fingerprints, utilising 2D and 3D structures. Additionally, for both drugs and targets we constructed Gaussian kernels based on the similarities of molecules' interaction profiles. We ensured positive-semidefinitness of all kernel matrices.

Results and Discussion The predictive performance of the Kronecker RLS was assessed using Pearson correlation. We performed a Leave-One-Out Cross Validation, where one drug-target pair at a time is being removed in the training phase. The best predictions were obtained for the feature set where we used extended targets' profile and a combination of two kernels for drugs: Gaussian kernel computed based on drugs' interaction profiles (D-pKi) and Tanimoto kernel calculated based on both 2D and 3D structural similarities of compounds (D-[2D+3D]). The average accuracy across compounds is equal to 0.77. Moreover, in case of drugs, we observed that utilising similarities of interaction profiles (D-pKi) is more beneficial than using their structural similarities (D-[2D+3D]). It might be because a minor structural difference between drugs can cause a dramatic change in their activity, and D-pKi kernel enables to capture such behaviour. Currently, we are in the process of experimentally validating the most promising predictions of originally unknown drug-kinase interactions.

[1] Metz J T, et al.: *Navigating the kinome*. Nature Chemical Biology 2011.

[2] Pahikkala T, et al.: *Toward more realistic drug-target interaction predictions.* Briefings in Bioinformatics 2014.

[3] Pahikkala T, et al.: *Efficient regularized least-squares algorithms for conditional ranking on relational data.* Machine Learning 2013.

Applications of Proteochemometrics – From Species Extrapolation to Cell Line Sensitivity Modelling.

Isidro Cortes-Ciriano, Gerard J.P. van Westen, Daniel S. Murrell, Eelke B. Lenselink, Andreas Bender and Therese E. Malliavin.

Proteochemometrics (PCM) is a computational technique to model the bioactivity of multiple ligands against multiple targets, *e.g.* proteins or cell lines, simultaneously. Therefore, PCM has enabled the exploration of the selectivity and promiscuity of ligands on different protein classes [1,2]. Indeed, the simultaneous inclusion of both chemical and target information permits the extra- and interpolation to predict the bioactivity of compounds on yet untested targets [3]. In this contribution, we will firstly show a methodological advance in the field [4], namely how Bayesian inference (Gaussian Processes) can be successfully applied in the context of PCM for (i) the determination of the applicability domain of a PCM model; (ii) the prediction of compounds bioactivity as well as the error estimation of the prediction; and (iii) the inclusion of the experimental uncertainty of bioactivity measurements during model training. Additionally, we will describe how PCM can be useful in medicinal chemistry to concomitantly optimize compounds selectivity and potency, in the context of two application scenarios, which are: (a) modelling isoform-selective cyclooxygenase inhibition; and (b) large-scale cancer cell line drug sensitivity prediction.

[1] GJP van Westen, JK Wegner, AP Ijzerman, HWT van Vlijmen, A Bender. *Proteochemometric Modeling as a Tool to Design Selective Compounds and for Extrapolating to Novel Targets.* Med. Chem. Commun. 2011, 2, 16-30.

[2] I Cortes-Ciriano, QU Ain, V Subramanian, EB Lenselink, O Mendez-Lucio, AP IJzerman, G Wohlfahrt, P Prusis, TE Malliavin, GJP van Westen, A Bender. *Polypharmacology Modelling Using Proteochemometrics (PCM): Recent Methodological Developments, Applications to Target Families, and Future Prospects*. In revision at Med. Chem. Commun.

[3] GJP van Westen, JK Wegner, P Geluykens, L Kwanten, I Vereycken, A Peeters, AP Ijzerman, HWT van Vlijmen, A Bender. *Which Compound to Select in Lead Optimization ? Prospectively Validated Proteochemometric Models Guide Preclinical Development.* PLoS ONE. 2011, 6, e27518.

[4] I Cortes-Ciriano, GJP van Westen, EB Lenselink, DS Murrell, A Bender, TE Malliavin. *Proteochemometric Modelling in a Bayesian framework.* Accepted at J. Cheminf. 2014.

Peak Finding on NGS Biological Replicates via Mixture Model Clustering

Mahmoud M. Ibrahim[1], Scott A. Lacadie[1] and Uwe Ohler[1]

[1] The Berlin Institute for Medical Systems Biology at the Max Delbrueck Center for Molecular Medicine Berlin-Buch

ABSTRACT

Although peak finding in NGS datasets has been addressed extensively, there is no consensus on how to analyze and process biological replicates. Furthermore, most peak finders do not focus on accurate determination of enrichment site widths and are not widely applicable to different types of datasets.

We developed JAMM (**J**oint **A**nalysis of NGS replicates via **M**ixture **M**odel clustering): a peak finder that can integrate information from biological replicates, determine enrichment site widths accurately and resolve neighboring narrow peaks. JAMM is a universal peak finder that is widely applicable to different types of datasets.

JAMM starts by selecting local windows that are enriched over background, followed by clustering the normalized extended-read counts in those windows into a peak cluster and noise cluster(s). Local clustering allows JAMM to adapt to peaks with different signal properties and to accurately determine their boundaries. Furthermore, using clustering as an approach for peak finding extends naturally to multivariate clustering, which is useful for integrating datasets that are highly correlated, such as biological replicates. We chose clustering via multivariate Gaussian mixture models, which allows for including information about the co-variance of the replicates. Finally, JAMM scores the reported peaks via the peak signal, represented by the per-position geometric mean of the replicates peak signals, and how it compares to background; thereby providing robust peak ranking.

We show that JAMM is amongst the best performing peak finders in terms of site detection specificity and in terms of accurate determination of enrichment site widths. In addition, JAMM's replicate integration improves peak finding resolution, sorting and peak finding specificity. JAMM is available for free and can run on Linux machines through the command line: http://code.google.com/p/jamm-peak-finder

EDEN: Experimental Design for parameter Estimation in a gene regulatory Network

Artémis Llamosi[1,4], Adel Mezine[1], Michèle Sebag[3], Véronique Letort[2], and Florence d'Alché-Buc[1,3]

[1] Informatique Biologie Intégrative et Systèmes Complexes (IBISC),
{artemis.llamosi@univ-paris-diderot.fr}
{amezine@ibisc.univ-evry.fr}
{florence.dalche@ibisc.univ-evry.fr}
Université d'Evry-Val d'Essonne, France
[2] Ecole Centrale Paris, 92295 Châtenay-Malabry cedex
{veronique.letort@ecp.fr}
[3] TAO, INRIA Saclay
Laboratoire de Recherche en Informatique (LRI), CNRS, Université Paris Sud, Orsay, France
{Michele.Sebag@lri.fr}
[4] Laboratoire Matière et Systèmes Complexes, Université Paris Diderot & CNRS, 75013 Paris, France
INRIA Paris-Rocquencourt, Rocquencourt, 78153 Le Chesnay, France

Abstract. Quantitative models are essential to study the dynamics of complex biological systems such as gene regulatory networks. Parametric estimation of such models is often hampered by the high cost of experiments and therefore the limited number of data. A too limited number of conditions in which observations are measured also raises the issue of practical non-identifiability for some of the parameters. In that context, a careful choice of experiments including perturbation such as knock-out or knock-down strongly makes the difference and allows to mitigate these non-identifiabilities. In this work, assuming that the biologist is given a budget to perform experiments, we address the problem of sequential experimental design in order to improve the accuracy of parameter estimates. We present a novel algorithm, called EDEN, which starts from an initial experimental dataset, then sequentially estimates the model parameters from partial and noisy observations and suggests a next experiment that could improve the quality of current estimation. Formulated as an active learning problem, the experimental design problem is modeled as a one-player game. An algorithm based on Upper Confidence Tree, combining Monte-Carlo tree search and multi-armed bandits, is proposed to explore the space of experiments sequences, to evaluate the utility of the most promising ones and to select the best ones.
Our approach is demonstrated on a realistic simulated gene regulatory network inspired from the international challenge DREAM7.

Keywords: active learning; experimental design; parameter estimation; Monte-Carlo tree search; upper confidence bounds for tree; ordinary differential equations; gene regulatory network

References

1. Browne, C. B., Powley, E., Whitehouse, D., Lucas , S. M., Cowling, P, I., Rohlfshagen, P., ... & Colton, S.: A survey of Monte-Carlo tree search methods. Intelligence and AI, 1(4), pp. 1–49 (2012)
2. Kocsis, L., and Szepesvári, C.: Bandit based Monte-Carlo planning. In: ECML-06. Number 4212 in LNCS, 282–293. Springer (2006)
3. Llamosi, A., Mezine, A., Sebag V., Letort V., d'Alché–Buc F. Experimental design in dynamical system identification: a bandit-based active learning approach, to appear in Proceedings of ECML/PKDD 2014, Nancy, France, (2014).
4. Mazur, J.: Bayesian Inference of Gene Regulatory Networks: From parameter estimation to experimental design. Ph.D. Dissertation, University of Heidelberg, Germany. (2012)
5. Quach, M., Brunel, N., and d'Alche Buc, F.: Estimating parameters and hidden variables in non-linear state-space models based on odes for biological networks inference. Bioinformatics 23:3209–3216 (2007)
6. Rolet, P., Sebag, M., and Teytaud, O.: Boosting active learning to optimality: A tractable Monte-Carlo, billiard-based algorithm. In Buntine, W. L., Grobelnik, M., Mladenic, D., and Shawe-Taylor, J., eds., ECML/PKDD (2),volume 5782 of LNCS, pp 302–317. Springer (2009)

Learning to combine Semantic Features for Neurolinglistic Decoding

Over the last decade, brain imaging has led to many works in various fields ranging from philosophy and psychology to neuroscience and artificial intelligence. In particular the analysis of functional Magnetic Resonance Imaging (fMRI) has become a primary focus of interest and in particular inferring some high-level knowledge from these has become a real and reachable challenge. A three dimensional fMRI image may contain approximately 20,000 voxels (volumetric pixels) that are activated with some predictable patterns when a human performs a particular cognitive task or when he is subject to a particular stimulus (e.g. visuals or words) [3]. In some ways one expects that it is possible to predict human thoughts from the brain activity.

Concerning visual stimuli, researchers have addressed both problems of inferring the fMRI image corresponding to a particular visual stimuli class [1], as well as inferring the visual stimuli class from the fMRI image [2]. We are interested here in the latter case. One key idea that has been exploited in the past is that different concepts are encoded by different brain regions and areas (i.e. voxels activated in specific areas). They hence investigated the possibility to define a predefined number of semantic binary features (e.g. is it mad made? Can it be held? ...) allowing to express a large number of classes (words), where each class corresponds to a specific joint setting of all these semantic features. Doing so their study proposed to learn a predictor from the fMRI image to this semantic feature space where the recognition of the visual concept class corresponding to a fMRI input is performed as a nearest neighbour search between the inferred semantic features vector and the semantic representation of all known classes. This approach was shown to enable performing up to some extent zero shot learning, i.e. recognizing inputs corresponding to new classes (provided the semantic representation of the new class is given).

This pioneering work had few limitations yet. It concerned a limited number of concepts and it worked much better with hand designed semantic feature space than with automatically derived features from corpus. We want to go further in this work and aim at designing state of the art results with automatically learned semantic features. We do so by first relying on recent advances on the learning of distributed representations of words (i.e. word embeddings) from Wikipedia corpus [4]. Although this approach yield promising results it did not succeed in reaching the accuracy of a manually designed semantic space. To go further we have investigated the combination of multiple embeddings derived from the corpus as well as and from linguistic resources (i.e., WordNet). The combination of multiple embeddings offers more capacity to the system, but at the same time it comes with more likely over-training. We also propose to efficiently deal with this problem using a multi-task lasso regularization framework where each task consists of predicting the semantic features of one particular semantic space. We show that this approach allows outperforming the use of a manual semantic space with the additional benefit to enable dealing with any new classes.

Keywords: Neurolinguistics, Brain Decoding, Word Embedding Representation, functional Magnetic Resonance Imaging, and Machine Learning.

References

[1] Mitchell, Tom M and Shinkareva, Svetlana V and Carlson, Andrew and Chang, Kai-Min and Malave, Vicente L and Mason, Robert A and Just, Marcel Adam: Predicting human brain activity associated with the meanings of nouns. Science Volume 320:5880 2008: 1191-1195.

[2] Mark Palatucci, Dean Pomerleau, Geoffrey E. Hinton, Tom M. Mitchell: Zero-shot Learning with Semantic Output Codes. NIPS 2009: 1410-1418.

[3] Pereira, Francisco and Botvinick, Matthew: A systematic approach to extracting semantic information from functional MRI data. Advances in Neural Information Processing Systems 2012: 2267-2275.

[4] Tomas Mikolov, Kai Chen, Greg Corrado, and Jeffrey Dean: Efficient Estimation of Word Representations in Vector Space. In Proceedings of Workshop at ICLR, 2013.

ISSCOR: An alignment-free method for comparative genomics analysis of synonymous codon correlations

Jan Radomski[1], Piotr Slominski[2], and Dariusz Plewczynski[1]

[1]Interdisciplinary Centre for Mathematical and Computational Modeling, University of Warsaw, Warsaw, Poland
[2]Centre de Génétique Moléculaire du CNRS & Université Pierre-et-Marie Curie (Paris-6), France

Motivation: Living organisms often have biased preferences for certain synonymous codons coding for the same amino acids. Despite extensive study, decisive rules that govern this bias have yet to be discovered. Postulated forces driving synonymous codon usage include: translational optimization, mRNA structural effects, protein composition and protein structure, gene expression levels, tRNA abundance differences between different genomes, tRNA optimization, mutation rates and mutation patterns. Local compositional bias and even gene length might play a role. It is also possible that regularities present in the sequential order of occurrence of synonymous codons (SC) contribute to synonymous codon usage; however, this latter factor has yet to be thoroughly investigated.

Methods: To analyze the role of sequential orders in synonymous codon usage, we devised a novel in silico method called ISSCOR (Intragenic, Stochastic Synonymous Codon Occurrence Replacement). This approach is based on two key principles: (i) that the Monte Carlo shuffling of synonymous codons must preserve the overall codon usage profile of each sequence and, simultaneously, (ii) that any such shuffling must not change amino acid order of a gene. Both principles make this technique particularly well suited for assessing possible fluctuations of, and outcomes from, codon bias between genes, groups of genes, or entire genomes.

Results: Here we apply the ISSCOR method to understand synonymous codon usage among different strains of influenza, a highly genetically diverse family of viruses for which variations in the patterns of codon usage might provide insight into viral evolution. We present examples of findings and results using the ISSCOR method to analyze complete viral genomes or very large collection of orthologous gene sets. These data suggest that many interesting biological mechanisms underlie the phenomenon of genetic code degeneracy.

Greedy cluster, a fast and sensitive method for grouping protein sequences

Fabio Rocha Jimenez Vieira[1] and Juliana Silva Bernardes[2]

[1]COPPE-UFRJ, Brazil; [2]LBCQ-UPMC, France

An important problem in computational biology is the automatic detection of protein families (groups of homologous sequences). Clustering these sequences into families is at the heart of most comparative studies dealing with protein evolution, structure and function. Many methods have been proposed for this task, but to the best of our knowledge their performance can vary when grouping protein families with different characteristics. Some methods achieve a good performance for grouping proteins with high sequence identity, but they fail on highly diverged proteins. Other methods are able to identify a small number of larger protein families, while others are specialized to cluster huge datasets. Here, we present a greedy approach that works well in all cases. The method, called Greedy Cluster (GC), can be divided into three steps. First, the centroid of clusters are chosen. To this end, we let all proteins be the centroid of a cluster. Second, proteins are added to a cluster if the similarity (distance to the centroid protein) is higher that a threshold. Naturally, incorrect solutions will be produced: identical clusters, subsets of the same clusters and clones (proteins belonging to more than one cluster). These inconsistencies are then addressed in the third step. To remove identical clusters we consider just the cluster with the highest density. A subset of a cluster is kept iff its density is greater than the original cluster. To eliminate clones, we compute their average similarity for each cluster. Next, we remove the clone with the worst average similarity until there is no more clones. Note that when a clone is removed average similarity must be recomputed. To evaluate our method, we used three manually curated datasets. The first is a standard collection of homologous proteins presenting high sequence similarity. The second is based on subsets from SCOP (Structural Classification of Proteins), and it aims to measure the performance of sequence clustering methods on distantly related homologous proteins. The third dataset contains larger protein families. Our results show that our method is robust to both sequence divergence and larger protein families, moreover it is fast enough to be used on very large datasets.

Pangenome-based strain level metagenomic profiling

Matthias Scholz[1], Doyle V. Ward[2], Thomas Tolio[1], Duy Tin Truong[1], Adrian Tett[1], Ardythe L. Morrow[3], and Nicola Segata[1]

[1]Centre for Integrative Biology, University of Trento, Italy
[2] Broad Institute, Cambridge, Massachusetts, United States
[3] Cincinnati Children Hospital Medical Center, Cincinnati, OH, United States

Metagenomics provides the opportunity to explore complex microbial populations in natural and human-associated ecosystems. When sequencing the whole genomic content of a sample (shotgun metagenome sequencing) we aim to obtain a complete picture of the microbial diversity in a specific environment. However, despite the richness of the available metagenomic datasets, current computational tools are based on computationally intensive assembly-based approaches or are limiting the resolution of the analysis to the species level. To increase the resolution up to strain level, provide the ability to characterize strain-specific gene repertoires, and potentially enable metagenomic-based epidemiological studies, we developed a novel Pangenome-based Phylogenomic Analysis (PanPhlAn) approach. Our assembly-free tool detects the presence or absence of each gene of the entire gene set of a species (pangenome) compiled using sequenced reference genomes, thereby capturing the individual gene set of the specific strain of the species of interest present in the sequenced microbiome. This enables both the identification of known organisms and the characterization of new strains of unknown gene composition. When metagenomic and meta-transcriptomic datasets are both available for the same specimen, PanPhlAn also provides gene-specific transcription rates of individual strains in a sample, thus exposing the "in-vivo" transcription activity not available with culture-dependent approaches. We validated PanPhlAn on several synthetic metagenomes obtaining very accurate strain reconstructions and applied it on 4 large metagenomic cohorts (∼10 Tb) showing the potentialities of the approach. We applied the novel approach on a large disease-associated cohort of pre-term infants assayed by both metagenomics and metatanscriptomics highlighting the in vivo transcription rates of important infant gut colonizers. This is in turn enabling the potential identification of strain-level genetic biomarkers associated with the diseases included in the cohort (necrotizing enterocolitis (NEC), sepsis, chorioamnionitis). PanPhlAn is distributed as an open source python-based tool.

Retrieval of Experiment

Sohan Seth, Ritabrata Dutta, and Samuel Kaski

Helsinki Institute for Information Technology, Finland

We study the task of retrieving relevant experiments given a query experiment. By experiment, we mean a collection of measurements from a set of 'covariates' and the associated 'outcomes'. While similar experiments can be retrieved by comparing available 'annotations', this approach ignores the valuable information available in the measurements themselves. To incorporate this information in the retrieval task, we suggest employing a retrieval metric that utilizes probabilistic models learned from the measurements. We argue that such a metric is a sensible measure of similarity between two experiments since it permits inclusion of experiment-specific prior knowledge. We discuss different approaches of evaluating this metric, and compare their pros and cons. We also present preliminary results on both simulated and real datasets showing the efficacy of such approach.

Metabolite identification through multiple kernel learning on fragmentation trees

Huibin Shen[1], Kai Dährkop[2], Sebastian Böecker[2], and Juho Rousu[1]

[1]Aalto University, Finland; [2]Friedrich Schiller University Jena, Germany

Metabolite identification from tandem mass spectrometric data is a key task in metabolomics. Various computational methods have been proposed for the identification of metabolites from tandem mass spectra. Fragmentation tree methods explore the space of possible ways in which the metabolite can fragment, and base the metabolite identification on scoring of these fragmentation trees. Machine learning methods have been used to map mass spectra to molecular fingerprints; predicted fingerprints, in turn, can be used to score candidate molecular structures.

We combine fragmentation tree computations with kernel-based machine learning to predict molecular fingerprints and identify molecular structures. We introduce a family of kernels capturing the similarity of fragmentation trees, and combine these kernels using recently proposed multiple kernel learning approaches. Experiments on two large reference datasets show that the new methods significantly improve molecular fingerprint prediction accuracy. These improvements result in better metabolite identification. Recent result on a 2825 compounds metlin MS/MS spectra dataset shows that 27% percent of the spectra can be identified correctly and 60% percent of the spectra can be identified within top 10 of the candidates list.

Disordered proteins in the eyes of a molecular chaperone

Magdalena Wawrzyniuk[1], Luca Ferrari[1], Madelon Maurice[2], and Stefan Rüdiger[1]

[1]Utrecht University, Netherlands; [2]UMC Utrecht, Netherlands

The Hsp90 family constitutes the most abundant cytoplasmic molecular chaperone system, which assists late stages of protein folding. Understanding substrate selectivity of the Hsp90 chaperone machine is crucial to understand protein folding in the cell. Recently, we obtained a structural model of Hsp90 in complex with one of its natural substrates, the Tau protein[1]. Remarkably, Tau is an intrinsically disordered protein. The Hsp90-Tau complex reveals how a disordered protein appears in the eyes of a chaperone. Based on this paradigmatic interaction, we set out to extract general themes of Hsp90 substrate recognition, which aims to provide a general mechanistic view on why and when a molecular chaperone can recognize intrinsically disordered proteins.

We developed an algorithm to identify stretches of similar properties in other disordered proteins. We employed a range of relevant parameters including hydrophobicity and charge patterns, and, involved several disorder probability predictors [2,3,4]. As first target, we focused on the instrinsically disordered scaffold proteins of the destruction complex of the Wnt signalling cascade. It contains well known Hsp90 interactors and numerous intrinsically disordered proteins. Based on this, we developed a bioinformatic tool for screening for potential Hsp90 binding sites among intrinsically disordered proteins. We are currently testing our predictions experimentally for a diverse set of substrates.

Reference

1. Karagöz GE, Duarte AM, Akoury E, Ippel H, Biernat J, Morán Luengo T, Radli M, Didenko T, Nordhues BA, Veprintsev DB, Dickey CA, Mandelkow E, Zweckstetter M, Boelens R, Madl T, Rüdiger SGD. Hsp90-Tau complex reveals molecular basis for specificity in chaperone action (2014) Cell 156, 963-74.
2. J. Cheng, M. Sweredoski, P. Baldi, Accurate Prediction of Protein Disordered Regions by Mining Protein Structure Data, Data Mining and Knowledge Discovery, vol. 11, no. 3, pp. 213-222, 2005.
3. Li, X., P. Romero, M. Rani, A. K. Dunker, and Z. Obradovic, Genome Informatics, 1999, 10:30-40.
4. Zsuzsanna Dosztányi, Veronika Csizmók, Péter Tompa and István Simon IUPred: web server for the prediction of intrinsically unstructured regions of proteins based on estimated energy content ; Bioinformatics (2005) 21, 3433-3434.

Sensitivity Analysis of Sinoatrial Node Model by Stochastic Simulation

Jianhao Xiong, and Mahesan Niranjan

July 31, 2014

The variability of heart rate is an important indicator of the physiological state of the heart. The rhythmic beating of the heart originates in the pacemaker cells of the sinoatrial node. Action potentials in the sinoatrial node cells are triggered by inward and outward flows of various ions via transmembrane ion channels. Non-linear regulation of ion channel opening and closing in a voltage dependent manner provides the feedback mechanism that enables sustained oscillations. Additionally, the autonomic nervous system has a regulatory role in the beating of the heart via sympathetic and parasympathetic inputs on ion channels.

Physical or mechanistic model of sinoatrial node, based on differential equations, governs the flow of ions through voltage regulated ion channels. As many as 13 ion channels can be involved in the gating of these ion flows [1] , and the properties and the interaction of these ion channels determine the firing rate and variability of the pacemaker at the cellular level.

In this work, we seek to explain clinically useful information about heart rate variability from cellular level physical models of sinoatrial node pacemaker cell function. We wish to explain correlates of observed variability in parameters variables of the physical models. Towards this goal, in this work, we present sensitivity analysis of model parameters in driving rhythmic behaviour of pacemaker cells. Our analysis, based on stochastic simulation, ranks the variables considered in terms of their sensitivities in regulating rhythmic behaviour. The contributions and the interactions of ion channels are presented to quantify the role of each ion channel.

References

[1] Z. Zhang and et al, *Mathematical models of action potentials in the periphery and center of the rabbit sinoatrial node*, Am J Physiol Heart Circ Physiol, 279(2000), pp. 397–421.