A LOGICS-BASED INTEGRATIVE APPROACH TO DECIPHER PUTATIVE REGULATORY RELATIONSHIPS INFERED FROM GENOMIC AND TRANSCRIPTOMIC DATA

ANDRÉS ARAVENA [1,2], DAMIEN EVEILLARD [3], ALEJANDRO MAASS [1,4], AND ANNE SIEGEL [5,2]

One usual question that researchers on gene expression ask is why two genes are co-expressed. In bacteria a first answer is given by the operon structure: two genes in the same operon are transcribed together and thus are co-expressed. A second approach is to look for shared regulators which potentially impact on the expression of the relevant genes. A well established way to determine these regulation relationships uses homology and/or position weight matrices to putatively determine regulator genes and their targets over the bacterial genomic sequence. These methods often yield thousands putative regulations, one order of magnitude over the number of experimentally validated regulations in model bacteria as *E.coli*.

In this work we present a method to filter those putative regulations by integrating information of different nature: putative regulatory relationships, operon predictions and co-expression evidence. The first two sources of information are based on sequence analysis, while the last one comes from microarray or other types of expression experiments.

This integration is implemented on the framework of Answer Set Programming, where facts derive from each information source, and their relationships are established as logical constraints. ASP is an efficient way to compactly describe this kind of criteria and explore the combinatorial space to determine the set of elements satisfying the constraints.

Our method first builds an oriented graph where each node is an operon and there is an edge between operons X and Y whenever there is a gene in X which putatively regulates one or more genes in Y. The weight of each edge is a discrete value among three (high, medium or low) or five levels, chosen in consideration to the confidence level of the putative regulation.

Then we integrate co-expression evidence, which can be stated using an index like mutual information, and use it as a constraint for a minimization problem: for each co-expressed pair of operons we look for the set of paths in the graph connecting a shared regulator to them, which minimize the total edge confidence level. The resulting graph has a reduced size respect to the inital prediction and includes regulation relationships which may not be explicit in the co-expression analysis.

In a test case using *E.coli* sequence and a set of microarray experiments in 240 conditions, the regulation graph reduced to 31.2% of the initial size, while keeping 55.3% of the regulations independently described on literature. An hypergeometric test shows that this ratio is significantly biased towards experimentally validated regulations when compared versus random selection (p-value $\approx 10^{-35}$).

In terms of predictions of regulations affecting any given operon, the average number of regulation relationships decreases from 5.7 to 1.9, while the precision with respect to known regulations increases from 24.2% to 42.9%.

This results allows any researcher to focus on a reduced number of regulations to validate. Also, it provides a small and manageable set of genes connected by regulation with the genes of his/her interest.

[1]CENTER FOR MATHEMATICAL MODELING (CNRS UMI 2807) AND CENTER FOR GENOME REGULATION, UNIVERSIDAD DE CHILE; [2]INRIA, CENTRE RENNES-BRETAGNE-ATLANTIQUE, PROJECT DYLISS, CAMPUS DE BEAULIEU, RENNES; [3]LINA (UMR 6241), UNIVERSITÉ DE NANTES, ECOLE DES MINES DE NANTES & CNRS, NANTES; [4]DEPARTMENT OF MATHEMATICAL ENGINEERING, UNIVERSIDAD DE CHILE; [5]CNRS UMR 6074, IRISA PROJECT DYLISS, UNIVERSITÉ DE RENNES 1.

Transcriptional regulatory network in *Arabidopsis thaliana* during response to single and combined stresses

Pankaj Barah[1], Naresh Doni Jayavelu[2], Simon Rasmussen[3], Henrik Bjørn Nielsen[3], John Mundy[4] and Atle M Bones[1]

[1]Department of Biology, Norwegian University of Science and Technology, Trondheim, Norway N-7491

[2]Department of Chemical Engineering, Norwegian University of Science and Technology, Trondheim, Norway N-7491

[3]Center for Biological Sequence Analysis, Department of Systems Biology, Technical University of Denmark, Lyngby, Denmark DK-2800

[4]Department of Biology, University of Copenhagen, Copenhagen, Denmark DK-2200

* To whom correspondence should be addressed.

Professor Atle M Bones
Cell Molecular and Genomics Group
Norwegian University of Science and Technology
Department of Biology
Hoegskoleringen 5
N-7491 Trondheim, Norway
Phone: +47-73598692
Fax: +47-73596100
Email: atle.bones@bio.ntnu.no

Keywords: regulatory networks, transcription factor, combine stress, transcriptomics, network component analysis,

ABSTRACT

Plants are sessile organism and have developed robust mechanisms to perceive external signals and subsequently evolved with complex networks of defence responses against different types of stresses [1]. Complex regulatory circuits consisting of transcriptional activators and repressors known as transcription factors (TFs), control the plant's defence transcriptome. Unfortunately, comprehensive genome scale understanding regarding activity of transcription factors (TFs) and their regulatory relationships with target genes are presently not yet available for the model plants like *Arabidopsis thaliana* (*A. thaliana*). Several computational algorithms have been developed to identify regulatory network modules and their condition-specific regulators from genome scale expression data. By employing a powerful systems biology approach- Network Component Analysis (NCA), we have constructed a differential stress regulatory network model in *A. thaliana* during 11 stress conditions and predicted regulatory connections that might be crucial during combined stress exposure in plants. For this purpose, we have used a unique and homogeneous gene expression dataset (total 207 arrays) from ten *A. thaliana* ecotypes [2], in response to 5 single and 6 combined stress conditions (59 unique stress experiments). Differential activities of stress specific and multiple-stress regulated TFs have been computed and analysed. Apart from retaining several previously known interactions (cross validated using AraNet, AthaMap), many novel interactions between key TFs and their respective target gens involved in the stress response in *A. thaliana* were suggested.

Ref.:

[1] Chawla, K., P. Barah, M. Kuiper and A. M. Bones (2011), Systems Biology: a promising tool to study abiotic stress responses, in Omics and plant abiotic stress tolerance, edited by N. Tuteja, S. S. Gill and R. Tuteja, Bentham Books (USA): 163-172, eISBN: 978-1-60805-058-1.

[2] Rasmussen, S., P. Barah, M. C. Suarez-Rodriguez, L. Gautier, A. M. Bones, H. B. Nielsen and J. Mundy "Transcriptome responses to combinations of stresses in Arabidopsis thaliana " Plant Physiol. 2013 (in press).

Sparse latent graphical models in high dimensional setting with application to genetics

Pariya Behrouzi[1], Frank Johannes[1], Ernst C. Wit[2]

February 15, 2013

[1] Groningen Bioinformatics Centre, University of Groningen, Groningen, Netherlands
p.behrouzi@rug.nl
[1] Groningen Bioinformatics Centre, University of Groningen, Groningen, Netherlands
frank@johanneslab.org
[2] Faculty of Mathematics and Natural Sciences, University of Groningen, Groningen, Netherlands
e.c.wit@rug.nl

We present a fast computational method for inferring a sparse graphical model for a large number of categorical variables. The area of high dimensional statistics deals with estimation in the "large p, small n" setting, where p and n correspond, respectively, to the number of random variables and sample size. Such high dimensional problems routinely arise in genetics where the number of measured genomic features exceeds the number of biological samples. As a motivating example we consider genetic inbreeding data in maize, where we aim to uncover the conditional dependence relationship between 273 DNA sequence markers. Moreover, we aim to construct the related gene association network. Graphical models provide a probabilistic tool to display analysis and visualize the dependence structures by drawing a graph describing the conditional dependencies between the variables.

Keywords: High-dimensional problems; Copula Gaussian graphical model; graphical lasso; latent variable; EM algorithm.

Gene regulatory network inference using a boosting algorithm and operator-valued kernels

Néhémy Lim *IBISC, EA 4526, Université d'Evry-Val d'Essonne, Evry, France and CEA, LIST, Gif-sur-Yvette, France*
Yasin Şenbabaoğlu *Department of Computational Medicine & Bioinformatics, University of Michigan, Ann Arbor, MI 48109-2218*
George Michailidis *Department of Statistics, University of Michigan, Ann Arbor, MI 48109-1107*
Florence d'Alché-Buc *INRIA-Saclay, LRI umr CNRS 8623, Université Paris Sud, France and IBISC, EA 4526 Université d'Évry-Val d'Essonne,Évry, France*

Abstract

Reverse engineering of gene regulatory networks remains a central challenge in computational systems biology, despite recent advances facilitated by benchmark *in silico* challenges that have aided in calibrating their performance. We consider wild-type time series data as inputs and assume that nonlinear dynamical models are particularly appropriate for the inference task. We introduce a novel nonlinear autoregressive model based on **operator-valued kernels** that simultaneously learns the model parameters, as well as the network structure.

A flexible boosting algorithm (OKVAR-Boost) that shares features from L_2-**boosting** and randomization-based algorithms is developed to perform the tasks of parameter learning and network inference for the proposed model. Specifically, at each boosting iteration, a regularized operator-valued kernel based vector autoregressive model (OKVAR) is trained on a random subnetwork. The final model consists of an ensemble of such models. The empirical estimation of the ensemble model's **Jacobian** matrix provides an estimation of the network structure.

This study makes a number of key contributions to the challenging problem of network inference based *solely* on time course data. It introduces a powerful network inference framework based on nonlinear autoregressive modeling and Jacobian estimation. The proposed framework is rich and flexible, employing penalized regression models that coupled with randomized search algorithms and features of L_2-boosting prove particularly effective as the extensive simulation results attest. The models employed require tuning of a number of parameters and we introduce a novel and generally applicable strategy that combines **bootstrapping with stability selection** to achieve this goal. The performance of the proposed algorithm is evaluated on a number of benchmark data sets from the DREAM3 challenge and then, on real datasets. The high quality results obtained strongly indicate that it outperforms existing approaches.

Gene Network inference for high-dimensional problems

Abdolreza Mohammadi[1], Ernst C. Wit[2]

February 15, 2013

[1] Johann Bernoulli Institute, University of Groningen, Groningen, Netherlands
a.mohammadi@rug.nl
[2] Johann Bernoulli Institute, University of Groningen, Groningen, Netherlands
e.c.wit@rug.nl

In gene Networks the objective of data-intensive studies is to infer the relationships between various actors under scrutiny. A graph is one possible way to describe these relationships. When data is obtained from noisy measurements of the nodes in the graph, then graphical models present an appealing and insightful way to describe graph-based dependencies between the random variables.

In this talk, first, we briefly present our recently developed framework for network determination [1] and a discussion of its performance for high-dimensional problems. The proposed method is easy to implement and computationally feasible for large graphical models. We illustrate the efficiency of the proposed methodology on simulated and real data sets. Besides, we have implemented the proposed methodology into an R-package, called BDgraph which is freely available online.

Keywords: Gene network inference; High-dimensional problems; Bayesian model selection; Birth-death process; Markov chain Monte Carlo; Gaussian graphical models.

References

[1] A. Mohammadi and E. C. Wit. (2013) **Gaussian graphical model determination based on MCMC inference**. arXiv:1210.5371.

CO-ASSOCIATION AND GENE NETWORK ANALYSIS FOR PIG INTRAMUSCULAR FATTY ACID COMPOSITION

Y. Ramayo-Caldas[1,2], A. Reverter[3], M. R. S. Fortes[4], M. Ballester[2], A. Esteve-Codina[5], J.L. Noguera[6], A.I. Fernández[7] M. Pérez-Enciso[1,8], J. M. Folch.[1]

Affiliations: [1]Departament de Ciencia Animal i dels Aliments, Facultat de Veterinaria, Universitat Autonoma de Barcelona, 08193, Bellaterra, Spain. [2]Centre de Recerca en Agrigenòmica (CRAG), Consorci CSIC-IRTA-UAB-UB, Campus UAB, Bellaterra, 08193, Spain. [3]Commonwealth Science and Industrial Research Organisation, division of Animal, Food and Health Sciences, Brisbane, Qld 4067, Australia. [4]The University of Queensland, Queensland Alliance for Agriculture and Food Innovation, Center for Animal Science, Gatton, Qld 4343, Australia.[5]Centre Nacional d'Anàlisi Genòmica (CNAG). Torre I. Baldiri Reixac 4. 08028, Barcelona, Spain. [6] Genètica i Millora Animal, IRTA Lleida, 25198, Lleida, Spain, [7] Departamento Mejora Genética Animal, SGIT-INIA, Ctra. Coruña Km7.5 28040, Madrid, Spain. [8]Institució Catalana de Recerca i Estudis Avançats , ICREA, Barcelona, Spain.
Email: yuliaxis.ramayo@uab.es

Pig (*Sus scrofa*) is both a source of food and an animal model for the study of metabolic diseases. Fatty acids (FA) are a major energy source, playing a relevant role as cellular signalling molecules in various metabolic pathways and its composition influences pork meat quality. In this study, the results from genome-wide association studies (GWAS) were exploited in an Association Weight Matrix (AWM) approach to predict gene networks related to intramuscular FA composition in pigs. The associations of single-nucleotide polymorphism (SNP) that were individually associated with a primary phenotype of interest were explored across 15 related traits. In the AWM matrix, rows represent genes or SNP and columns represent traits. Each {i, j} cell value corresponds to the z-score normalized additive effect of the ith gene (via its neighbouring SNP) on the jth trait. Columnwise, the AWM recovered the genetic correlations estimated via pedigree-based restricted maximum-likelihood methods. Gene-gene or gene-SNP interactions were predicted using pairwise correlation from the standardized additive SNP effects across AWM rows. Gene Ontology (GO) and Pathway enrichment analysis were performed to study the predicted gene network and an overrepresentation for GO terms and pathways related to lipid metabolism was observed. To identify potential regulators, we focussed on the transcription factors (TF) found in the gene network. After computing all possible combinations of TF trios we identified three TF which are network hubs: *Nuclear receptor coactivator 2* (*NCOA2*), *E1A binding protein p300* (*EP300*) and *four and a half LIM domains 2* (*FHL2*). All three key TF have Transcription Factors Binding Sites in their promoter region for some well know TF that are considered as important regulators of lipid and carbohydrate metabolism. According String database, experimental data confirmed that protein-protein interaction exist among the three identified TF and those master regulators of lipid and carbohydrate metabolism. Additionally, 39 of the AWM-predicted target genes have been recently reported in two large-scale meta-analysis studies for plasma lipids in humans. Interestingly, many of these genes, including the three key TF, would be missed by the traditional single-trait GWAS. As noted before and confirmed by this study, AWM points to new candidate genes, TF and gene interactions via exploring SNP co-associations across multiples traits beyond the one-dimensional approach for identifying genes affecting specific traits. In conclusion, our results suggest a cooperative role for the three TF in the transcriptional regulation of Intramuscular FA composition and the control of energy homeostasis in pigs. In addition, we provide additional evidence supporting the pig as an animal model for the study of metabolic diseases in humans.

Joint estimation of causal effects from observational and intervention gene expression data

Andrea Rau[1], Florence Jaffrézic[1] and Grégory Nuel[2]

[1] INRA, UMR 1313 GABI, Jouy-en-Josas, France

[2] MAP5, UMR CNRS 8145, University Paris Descartes, Paris, France

Methodological development for the inference of gene regulatory networks from transcriptomic data is an active and important research area. Several approaches have recently been proposed to infer relationships among genes from observational steady-state expression data alone, primarily based on the use of graphical Gaussian models [1]. However, these methods rely on the estimation of partial correlations and are only able to provide undirected graphs that cannot highlight causal relationships among genes. A major upcoming challenge is hence to jointly analyze observational transcriptomic data with intervention data obtained by performing knock-out or knock-down experiments in order to uncover causal gene regulatory relationships.

Two methods have recently been proposed for causal gene network inference from observational and intervention data. Pinna *et al.* [4] propose a direct, naive comparison of observed expression values to expression under each intervention through the calculation of a standardized deviation matrix. In a more sophisticated approach, Maathuis *et al.* [3] use the PC algorithm [2] to produce partially directed networks using observational data, and intervention data are subsequently used to direct edges otherwise left undirected. When intervention data are unavailable, the PC algorithm arbitrarily chooses the direction of edges to avoid specific structures in the graph, such as cycles. Although this method has been shown to enable prediction of strong causal effects from observational data alone [5], it has a major drawback: the skeleton of the graph is fixed after estimation using the observational data, and cannot be changed based on contributions from the intervention data.

The aim of this paper is to propose a MCMC algorithm to jointly infer causal gene networks from observational and intervention data in the context of Gaussian Bayesian networks. In our approach, the structure of the graph can be modified based on the additional information brought by the intervention data. A classical Metropolis-Hastings algorithm was found to perform poorly for this purpose and was much improved by an empirical Bayes approach with a maximization over parameters for a given graph structure. We compare our proposed method to those of Pinna *et al.* and Maathuis *et al.* on simulated and real data from the DREAM (Dialogue for Reverse Engineering Assessments and Methods) 2007 challenge. Our simulation study confirms the difficulty in accurately estimating causal relationships from observational data alone, while the addition of intervention data allowed at least a partial recovery of these relationships. We also demonstrated that multi-factorial intervention experiments could provide more information than mono-factorial ones, thus raising the question of optimal experimental knock-out designs for future work.

References

[1] J.H. Friedman, T. Hastie, R. Tibshirani (2008) Sparse inverse covariance estimation with the graphical lasso. *Biostatistics*, 9(3):432-441.

[2] M. Kalisch, M. Mächler, D. Colombo, M.H. Maathuis, P. Bühlmann (2012) Causal Inference Using Graphical Models with the R Package pcalg. *Journal of Statistical Software*, 47(11):1-26.

[3] M.H. Maathuis, M. Kalisch, P. Bühlmann (2009) Estimating high-dimensional intervention effects from observational data. *Annals of Statistics*, 37:3133-3164.

[4] A. Pinna, N. Soranzo, A. de la Fuente (2010) From knockouts to networks: establishing direct cause-effect relationships through graph analysis. *PloS one*, 5(10):e12912.

[5] M.H. Maathuis, D. Colombo, M. Kalisch, P. Bühlmann (2010). Predicting causal effects in large-scale systems from observational data. *Nature Methods*, 7(4):247-248.

A general method for inferring causal relationships between associated phenotypes using both phenotypic and QTL information

Huange Wang and Fred van Eeuwijk
Biometris, Wageningen University, 6708 PB Wageningen, the Netherlands

Motivation: In addition to the analysis of genotype-phenotype relationships, mapping interactions between phenotypes also provides structural insight into the functional mechanism of biological systems. Various methods have been proposed to reconstruct directed phenotype networks. A recent interesting proposal, the QTL-directed dependency graph (QDG) approach, uses QTL information on phenotypes to infer causal directions for edges in an undirected phenotype network. A prerequisite for this approach is that at least one QTL has been identified for each trait studied. In practice, however, this prerequisite is often not met due to factors such as limited sample size, weak QTL effects and measurement noise.
Results: We developed a general method to infer causal directions for edges in a large-scale undirected phenotype network, using both the relevant phenotypic interactions and the detected QTLs. Our method does not require QTLs for each and every trait. We evaluated and compared the performance of our method with the benchmark QDG algorithm via simulations. Results show that our method is applicable to general cases and leads to more accurate overall orientations. Finally, we illustrated our method with a real example involving metabolic and QTL data in ripe tomato fruits.

What can biological networks tell us about the fate of duplicate genes in *Arabidopsis thaliana*?

Justin Whalley*, Etiene Birméle & Carène Rizzon

Laboratoire Statistique et Génome

UMR8071 CNRS, 23 bvd de France, Evry 91037, France

Background: Gene duplication is readily accepted as a primary mechanism for generating organismal complexity. However, the mechanisms responsible for the maintenance of duplicated genes at the genome scale are still poorly understood. Analysis of biological networks can help us to understand better which evolutionary forces are acting on duplicated genes, as their interacting context is taken into account. Using protein-protein interaction networks and gene regulatory networks, inferred from gene expression data, we look to investigate the functionalisation of duplicate genes in *Arabidopsis thaliana*.

Description: We used the clustering algorithm Walktrap[1] to define duplicate genes into families from a set of potential paralogous genes (found using BLAST). Making use of the scale free nature of protein-protein interaction networks we could investigate what effect duplicating a node (gene) has on the network, using such metrics as the Jaccard index, and the centrality and connectivity of nodes within the network. The gene-regulatory networks for *Arabidopsis thaliana* were inferred from gene expression data using SIMoNe.[2] Within these sparse networks, the length of the shortest path between duplicate genes gives an indication to what functionalisation the duplicate gene pairs under went.

Conclusions: Using the *Arabidopsis thaliana* protein-protein interaction network, our study showed gene function is linked to family size. This is in agreement with previous studies in vertebrates and invertebrates genomes. While the gene regulatory networks, implied that the functionalisation of duplicate genes is in part dependent on the type of duplication. Of which we find precedence in previous studies investigating gene expression data for *Arabidopsis thaliana* under different experimental conditions. Using biological networks to investigate the genome, has helped us to understand which genes can undergo duplication without overly affecting the protein-protein interaction network. Whilst in turn, gene regulatory networks can look into the functionalisation of genes after duplication.

[1]Pascal Pons and Matthieu Latapy. Computing communities in large networks using random walks.
Computer and Information Sciences – ISCIS 2005, volume 3733 of *Lecture Notes in Computer Science*, pages 284–293. Springer Berlin / Heidelberg, 2005.

[2]Julien Chiquet, Alexander Smith, Gilles Grasseau, Catherine Matias, and Christophe Ambroise. Simone: Statistical inference for modular networks.
Bioinformatics, 25(3):417–8, Feb 2009.

* Corresponding author – justin.whalley@genopole.cnrs.fr