GENE REGULATION AND TRANSCRIPTOMICS

Chairs: Jaak Vilo and Zohar Yakhini

E-1. TFactS: a tool to predict transcription factor regulation from gene expression data

Essaghir A (1,), Pachikian BD (2), Delzenne NM (2), Toffalini F (1), van Helden J (3), Demoulin JB (1)*

Deciphering transcription factor networks from microarray data remains difficult. This study presents a simple method to infer the regulation of transcription factors from microarray data based on their well-characterized target genes.

Materials and Methods

We generated a catalog containing 343 transcription factors associated with 2720 target genes and 6401 experimentally validated regulations. When it was available, a distinction between transcriptional activation and inhibition was included for each regulation. Next, we built a tool (TFactS) that compares submitted gene lists with target genes in the catalog to detect regulated transcription factors. TFactS was validated with published lists of regulated genes in various models and compared to tools based on in silico promoter analysis and was used to screen cancer gene expression data.

Results

On cancer data, TFactS showed the regulation of SOX10, MITF and JUN in melanomas. We then performed microarray experiments comparing gene expression response of human fibroblasts stimulated by different growth factors. TFactS predicted the specific activation of Signal transducer and activator of transcription factors by PDGF-BB, which was confirmed experimentally. We also compared in vivo gene expression in the liver of mice subjected to an omega3 fatty acid deficient diet. TFactS predicted an activation of LXR and SREBP which was confirmed experimentally.

Discussion

Altogether, we show here that comparing groups of genes showing a response in microarray data with experimental target gene signatures is an efficient way to predict the regulation of well-characterized trans-acting factors. As a proof of concept, we validated a simple tool (TFactS) that combines Fisher's test and a curated target catalogs, which do or do not take into account the sign of the regulations. We also suggest that TFactS may contribute to the functional analysis of cancer microarray data.

URL

http://www.tfacts.org

Presenting Author

Ahmed Essaghir (ahmed.essaghir@uclouvain.be)
Université Catholique de Louvain

Author Affiliations

1- de Duve Institute, Université Catholique de Louvain (UCL), Brussels. 2- Metabolism and Nutrition Research Group, Louvain Drug Research Institute, UCL, Brussels. 3- Bioinformatique des Génomes et des Réseaux (BIGRe), Université Libre de Bruxelles.

Acknowledgements

Wallonie-Bruxelles International scholarship.

E-2. Time- and dose-dependent effects of cigarette smoke on monocytic cell transcriptome

Poussin C (1,), Lietz M (2), Schlage W (2), Lebrun S (1), Hoeng J (1), Peitsch M (1)*

Atherosclerosis is a chronic inflammatory process characterized by a series of biological processes resulting in plaque formation in the vascular wall. The monocyte/macrophage might be an important key player in the initiation and progression of plaque formation. It has been shown that exposure to cigarette smoke (CS) increases the adhesion of monocytes to endothelial cells in vitro, a critical step in the initiation of atherosclerosis. The purpose of this study was to investigate the molecular mechanisms underlying the effect of CS exposure on monocytic cells over time.

Materials and Methods

Human Mono Mac 6 cells were treated with different doses (0.045 and 0.09 puff/ml) of CS that had been bubbled through phosphate buffered saline (PBS) or with control PBS for 30, 60, and 120 minutes. RNA from different cell samples was extracted for further transcriptomic analysis using the Human Genome U133 Plus 2.0 Affymetrix® gene expression platform. The statistical analysis, including a multiple linear model approach performed with the limma package (Smyth GK,2005), identified 502 probesets (~428 genes) with a significant interaction (FDR≤0.01) between the time and dose variables.

Results

Genes involved in transcription and transcription regulation, cell cycle, endoplasmic reticulum- and oxidative-stress response were over-represented in the list of differentially regulated genes. Indeed, numerous genes involved in the Nrf2-mediated oxidative stress response and in the p53-signaling pathway were found to be coordinately activated in a dose- and time-dependent manner. In fact, the expression of the cycle-dependent kinase inhibitor p21 (CDKN1A) gene was found to be dose- and time-dependently increased, whereas the expression of some cyclins, targets of p21, was repressed by CS.

Discussion

Interestingly, CDKN1A is an important negative regulator of cell cycle and has been shown to be involved in monocytic cell differentiation to macrophages and to protect against atherosclerosis when knocked out in apoE-/- mice. Overall, the results indicate that upon CS exposure, monocytic cells undergo stress responses and transcriptional gene expression regulation in a dose- and time-dependent manner, which may drive these cells to acquire new biological functionalities that may be involved in the development of atherosclerosis.

Presenting Author

Carine Poussin (carine.poussin@pmintl.com)
Philip Morris International R&D Neuchâtel

Author Affiliations

(1) Philip Morris International R&D, Philip Morris Products S.A., Neuchâtel, Switzerland. (2) Philip Morris International R&D, Philip Morris Research Laboratories GmbH, Cologne, Germany.

E-3. Combined Monte Carlo and dynamical modelling of gene regulation enables multiple parameter estimations and hypothesis generation

Herman D (1,), Thomas CM (2), Stekel DJ (3)*

KorA and KorB are global regulators of the broad host range plasmid RK2 that can either separately or cooperatively autorepress the central control operon (cco). Separate repression has been known as partial. However their mechanism and strength have not been examined. Better understanding of the cco control and a hypothesis formulation required a model generation and estimation of protein synthesis and monomerization rates. Our interest was also in the model sensitivity analyses and their consistency with information provided by the parameter estimation algorithm.

Materials and Methods

The gene regulation of the cco was mathematically modeled with a deterministic approach and the steady state of the growth phase was considered. Five models, which varied by different scenarios of the mechanism of partial repression, were examined. In order to estimate unknown parameters, the Metropolis-Hastings algorithm, based on a Markov Chain Monte Carlo method, was applied. This allowed us to include both known and unknown parameters into the model fitting by choosing log normal or flat priors as appropriate. For model sensitivity analyses, the Metabolic Control Analyses was implemented.

Results

The stochastic sampling of the parameter has resulted in multimodal posterior distributions, suggesting a number of hypotheses about the behavior of the system. Updated posteriors of monomerization rates have been spread over a broad value range of 50 orders of magnitude, implying an insignificance of these parameters for the model. This conclusion is consistent with the results from the MCA. A ratio of protein abundances for the examined models and the model without repression and experimental results of protein abundances for the unrepressed operon might be able to eliminate a real model.

Discussion

The reported results have demonstrated the general utility of the Monte Carlo Markov Chain (MCMC) approach to systems biology modeling. Additionally, the consistency in some information from MCMC and Metabolic Control Analyses calculations is significant. Moreover, based on our results an experimental verification we might be able to eliminate a model for the mechanism of partial repression of the central control operon on RK2 plasmid.

Presenting Author

Dorota Herman (dxh885@bham.ac.uk)
University of Birmingham

Author Affiliations

(1) Center for Systems Biology, School of Biosciences, The University of Birmingham, B15 2TT,United Kingdom (2) School of Biosciences, The University of Birmingham, B15 2TT,United Kingdom (3) Integrative Systems Biology, School of Biosciences, The University of Nottingham, LE12 5RD, United Kingdom

Acknowledgements

The Darwin Trust of Edinburgh

E-4. Module-based comparative gene expression analysis: evolutionary conserved coexpression in Bacillus subtilis and Escherichia coli

Zarrineh P (1,), Fierro C (2), Sánchez-Rodríguez A (2), De Moor B (1), Marchal K (2)*

Increasingly large scale expression compendia for different species are becoming available. By exploiting the modularity of the coexpression network, these compendia can be used to identify pathways or processes for which the expression behavior is conserved between different species. However, comparing module networks across species is not trivial as the definition of a biologically true module is not a fixed one, but depends on the distance threshold that defines the degree of coexpression with the modules.

Materials and Methods

We developed a cross-species cocluster approach COMODO (COnserved MODules across Organisms). COMODO uses as input microarray data and homology mapping among genes and provides as output pairs of conserved modules. COMODO searches for conserved coexpression modules in heterogeneous microarray expression compendia. Module sizes are selected as such that they maximize the number of orthologs that link up both modules relative to the total number of genes in each module that contribute to a conserved module pair.

Results

We applied our methodology to study coexpression conservation between Escherichia coli and Bacillus subtilis. We identified several pathways and processes for which transcriptional coexpression has been conserved over the long evolutionary distance that separates both prokaryotic organisms. As was expected, operons seemed to be one of the major mechanisms to guarantee coexpression behavior amongst genes, not only within a species, but also across species. The mechanism by which genes were transcriptionally coregulated seemed to be much less conserved than their coexpression behavior itself.

Discussion

We developed a method to search for conserved modules between two species. Applying this method allowed detecting both processes for which the expression behavior is conserved over the wide evolutionary distance that separates E. coli from B. subtilis or where subparts of the processes differ in expression As it relies on many-to-many homology mapping, instead of using bidirectional reciprocal best blast hit it offers the possibility to search for functionally true orthology relations.

Presenting Author

Peyman Zarrineh (Peyman.Zarrineh@esat.kuleuven.be)
Katholieke Universiteit Leuven

Author Affiliations

(1) Department of Electrical Engineering, Katholieke Universiteit Leuven, Kasteelpark Arenberg 10, 3001 Leuven, Belgium.
(2) Department of Microbial and Molecular Systems, Katholieke Universiteit Leuven, Kasteelpark Arenberg 20, 3001 Leuven, Belgium.

Acknowledgements

This work is supported by: 1) KUL: GOA AMBioRICS, GOA/08/011, CoE EF/05/007, SymBioSys; CREA/08/023 2) IWT: SBO-BioFrame; 3) IUAP P6/25 (BioMaGNet); 4) FWO IOK-B9725-G.0329.09; 5) ZKB8933/CREA/08/023/ BOF, 6) HFSP-RGY0079/2007C.

E-5. Algorithm-driven artifacts in median polish summarization of microarray data

Giorgi F-M (1,), Bolger A(1), Lohse M(1), Usadel B(1)*

High-throughput measurement of transcript intensities using Affymetrix type oligonucleotide microarrays has produced a massive quantity of data during the last decade. Different preprocessing techniques exist to convert the raw signal intensities measured by these chips into gene expression estimates. Although these techniques have been widely benchmarked for differential gene expression analysis, there are only few examples where their performance has been assessed in respect to other contexts like sample clustering.

Materials and Methods

We assessed overall correlation between several thousand Arabidopsis thaliana ATH1 microarrays available from Gene Expression Omnibus database using RMA, GCRMA and MAS5 preprocessing methods. We permuted the arrays to have a null-information scenario for correlation. We implemented a modified RMA version (tRMA) in R using this corrected summarization method. Finally, we tested clustering performance of our method using publicly available Arabidopsis and Human datasets.

Results

We benchmark the most used normalization procedures (MAS5, RMA and GCRMA) in the context of inter-array correlation analysis, confirming and extending the finding that RMA and GCRMA overestimate sample similarity. This artificial behaviour is particularly strong in internally inconsistent, noise-driven probesets. We offer a mathematical explanation for the artifact behaviour and relate this to an analysis of probesets' features. Our proposed solution, tRMA, modifies the summarization step of RMA/GCRMA and provides improved clustering performance in biological datasets.

Discussion

Our finding proposes that one of the most relevant advantages of RMA and GCRMA in microarray preprocessing benchmarks, the low variance across replicates, is partially the result of an artificial inter-array correlation. While this doesn't seem to affect differential gene expression analyses, it can't be ignored where unbiased measurements are needed, like sample clustering. We propose the tRMA method to overcome this issue, while retaining most of the features of the original RMA procedure.

Presenting Author

Federico M. Giorgi (giorgi@mpimp-golm.mpg.de)
Max Planck Institute for Molecular Plant Physiology

Author Affiliations

(1) Max Planck Institute for Molecular Plant Physiology, Am Muehlenberg 1, 14476 Golm Germany

Acknowledgements

Max Planck Society

E-6. In silico study of the regulation of sRNAs in Escherichia coli

Ishchukov I (*), Ryan D, Zarrineh P, Cloots L, Thijs I, Engelen K, Marchal K

In bacteria a large part of the sRNAs regulates gene expression by pairing to mRNAs and affecting the mRNA stability and/or translation. This sRNA network and its impact on bacterial regulation is still unknown: it is not clear how many targets are affected by one particular sRNA, nor how many sRNAs can regulate a single gene. Moreover, except for the fact that some known TFs regulate sRNAs, it is still unclear how the sRNA network is interwoven with that of the known part of the transcriptional network.

Materials and Methods

In this work we aimed at updating the known Escherichia coli transcriptional network with the sRNA regulatory network. This includes predicting the regulation of the sRNAs and extending sRNA regulons of with novel targets. sRNA regulation was studied using de novo motif detection based on phylogenetic footprinting. Intergenic sequences of orthologous sRNA families in different related gamma proteobacteria were subjected to the PhyloGibbs (v. 1.2). Parameters were optimized as described in Storms et al. For target prediction we combined the intaRNA and targetRNA.

Results

We updated E. coli sRNA interaction network and integrate it with the existing transcriptional network.

Discussion

We further validated the TF-motif assignments in our sRNA dataset by using a strategy based on GO enrichment, assuming that if the target genes of respectively the TF predicted to regulate the corresponding sRNA, and the sRNA itself are involved in similar functions, the assignment of the cognate TF to the sRNA should be more reliable.

Presenting Author

Ivan Ishchukov (ivan.ishchukov@student.kuleuven.be)
K U Leuven

Author Affiliations

K U Leuven

Acknowledgements

OE-grant KULeuven

E-7. Systems biology analysis at VIB

Ren X-Y, Bonnet E, Joshi A, Maere S, Michoel T, Van Parys T, Vermeirssen V, Van de Peer Y*

Lost in regulation networks? Want to find regulators in your expression data? Want to visualize the regulation networks? We have solutions for you!

Materials and Methods

We have some well benchmarked softwares that are specialized in inferring gene regulation networks by integrative modeling of large amounts of high-throughput molecular data using systems biology approaches.

Results

LeMoNe: a software package for inferring regulatory modules and predict their condition-dependent regulators from large-scale compendia of gene expression data. ENIGMA: a software tool to extract gene expression modules from perturbational microarray expression data. Pathicular: a Cytoscape plugin for studying the cellular response to genetic perturbations. MINT: a Cytoscape plugin for the identification of functional modules in integrated networks.

Discussion

If studying regulatory networks from systems biology perspective is all what you want, just try our tools! You can either download them from our website (http://bioinformatics.psb.ugent.be/software), or alternatively, you can bring your expression data to us for the analysis. The results will aid you greatly in the biological interpretations and in the assessment of crosstalk between biological processes, as well as speed up your possible biomarker discovery.

URL

http://bioinformatics.psb.ugent.be/software

Presenting Author

Xin-Ying Ren (xinying.ren@psb.vib-ugent.be)
Plant Systems Biology, VIB

Author Affiliations

Bioinformatics and Systems Biology, Plant Systems Biology, VIB

Acknowledgements

We acknowledge the support of Ghent University (Multidisciplinary Research Partnership “Bioinformatics: from nucleotides to networks”).

E-8. A computational framework for automated analysis of nonlinear dynamics of gene regulatory networks

Ironi L(*), Panzeri L

A specific class of ODEs adequately describes the essential features of the complex dynamics of GRNs. But, the qualitative behaviour of a GRN depends crucially on the parameter values, and bifurcation phenomena may occur when parameters pass critical values. Then, numerical investigations of large GRNs are nontrivial and time-consuming, and might also be impracticable as quantitative information on the biochemical reactions underlying regulatory interactions, and on parameters are often unknown. Thus, the need for an automated symbolic procedure to reveal the full range of network dynamics.

Materials and Methods

We assume threshold dependent regulation, i.e. only effective above or below a certain threshold. Switch-like behaviours across variable thresholds are properly modelled by steep sigmoid functions, and the ODES model linear and nonlinear dynamics occurring at different time scales. Unlike other tools, we consider self-regulation and directly deals with continuous response functions. The algorithm integrates QR methods with singular perturbation analysis: the former ones allow us to cope with incomplete knowledge, and the latter one to deal with processes that occur at different time scales.

Results

The algorithm works under the assumptions above and that any two genes are never regulated at the same threshold by a certain variable. Under these biologically reasonable assumptions, we have worked out the singular perturbation approach to establish sound rules, computationally tractable, that determine the trajectories as the sequence of phase space regions through which the system passes. When suitable conditions are fulfilled, the simulation outcomes do capture the network dynamical properties dependent on the model structure and invariant for ranges of model parameter values.

Discussion

The tool proposed is applicable to predict the full range of dynamics of GRNs that fulfill the given assumptions. Thus, it provides grounds for dry experiments in both Systems and Synthetic Biology. In the former context, it is useful to formulate new hypotheses that explain data previously unobserved or not interpretable by the current knowledge; in the latter one to preliminarly dry design of a synthetic regulatory network. In particular, we are applying the computational tool to explore the mechanisms underlying sex determination in Ceratitis capitata, and motility in Bacillus subtilis.

Presenting Author

Liliana Ironi (ironi@imati.cnr.it)
IMATI-CNR

Author Affiliations

IMATI-CNR, Pavia

E-9. Exploiting ESTs to infer plant alternative splicing

Potenza E (1,), Cestaro A (1), Fontana P (1), Bianco L (1), Velasco R (1)*

The aim of this work is to devise a method to estimate and analyze alternative splicing events from cDNAs information. In the past, AS was considered a minor mechanism in plants but recent studies show that AS accounts for 20 to 40% of the whole plant transcriptome. Given the increase of available expression data, we focused on plant genomes. The major goal of this work is the development and validation of a piece of software for predicting AS events from expressed sequences aligned to a genome.

Materials and Methods

FindAS was completely developed in Java, due to its capability of being cross-platform.The algorithm first reads alignment data from a gff3 format, then it performs two clustering processes. The first one is used to group together all the cDNAs sorted by genome start positions, while the second one is aimed at identifying “alignment blocks” inside the clusters created by the first process. The ratio between the coverage distribution of blocks and clusters is afterwards used to identify putative AS events. Finally, findAS predicts all possible transcripts by traversing a direct acyclic graph.

Results

Validation of findAS was performed on Arabidopsis thaliana,a comparison was made with ASIP database and the software has been assessed by using TAIR predictions.Predictions were compared with several dataset of TAIR annotation. FindAS has always proved to be slightly more specific and sensitive than ASIP, when compared on the same alignment.An interesting result is the number of predicted genes with alternative splicing event. If we consider intron retention and exon skipping, ASIP predicts 4790 genes while findAS 4439 but TAIR 8 contains only 4325 AS genes with alternative donor/acceptor too.

Discussion

The main problem for findAS, as well as for all the other AS predictors, is the low specificity; this is probably due to the lack of complete AS datasets suitable for a comprehensive analysis that can also include transcripts targeted by mRNA NMD or RUST surveillance pathway. It is likely that RNA-seq techniques will increase the number of available data and this will constitute a benefit for tools like findAS. As future work, we plan to develop findAS further to use deep sequencing information for improving the discrimination among false and non-coding alternative spliced transcripts.

Presenting Author

Emilio Potenza (emilio.potenza@iasma.it)
Fondazione Edmund Mach - Istituto Agrario di San Michele all'Adige

Author Affiliations

(1) Fondazione Edmund Mach - Istituto Agrario di San Michele all'Adige

Acknowledgements

Provincia Autonoma di Trento

E-10. Inferring gene regulatory networks to identify conserved regulation between Arabidopsis and Populus

Netotea S (1,), Hvidsten TR (1)*

Trees are a major source of renewable raw material in terms of cellulose and lignin biosynthesis, the two most common organic compounds on Earth. Populus is the only tree with a fully sequenced genome, and is thus a model organism for trees. The challenge now lies in taking advantage of the growing amount of 'omics' data, to describe the regulation of growth and differentiation in trees. The present study aims to develop efficient methods for inferring the gene regulatory networks of trees and to identify their unique regulatory mechanism by network alignment to another plant species.

Materials and Methods

We employ microarray gene expression data derived from 1024 hybridization experiments using spotted cDNA arrays, mapped to the second generation of Populus gene models. We infer networks of significant gene regulations based on several statistical dependency scores, including correlation, mutual information and a context-relevant approach that takes into consideration the local network neighborhood. We then compare the inferred network to Arabidopsis using several local and global network alignment scoring metrics. The alignment is based on both network topology and network function.

Results

The performance of various gene regulatory network inference methods is quantified by using multiple regression based on the network topology to predict the expression profiles of each gene in a cross validation framework. We align the Populus networks to the Arabidopsis network, present global alignment scores, investigate the correlation of various node centrality measures among the two networks and establish the high scoring network alignment motifs. We identify conserved regulation of cellular and tissue function, and focus on unique features related to trees.

Discussion

We combine function and topology in a method of local and global regulatory network alignment to compare the transcriptional programs of Populus and Arabidopsis. Our current goal is to develop methods to predict functional interaction that can complement traditional sequence based studies in plant biology. The next step will be integrating the transcriptional data with protein and metabolic data.

Presenting Author

Sergiu Netotea (sergiu.netotea@plantphys.umu.se)
Umea Plant Science Centre, Umea University

Author Affiliations

1 Bioinformatics lab, Umeå Plant Science Centre, Dep. of Plant Physiology, Umeå University, Sweden

Acknowledgements

The Swedish Research Council (VR) and The Swedish Governmental Agency for Innovation Systems (VINNOVA).

E-11. BayesPI: a new model to study protein-DNA interactions; a case study of condition-specific protein binding parameters for yeast transcription factors

Junbai Wang (1,), Morigen (2)*

The application of statistical-mechanical theory in computational prediction of TF binding sites, using ChIP-chip dataset, has generated numerous new algorithms. Most of the early developed methods are based on an assumption of low protein concentration, where the protein binding probability in the models is approximated by Maxwell-Boltzmann probability. For a full biophysical modeling of protein-DNA interaction, a term called chemical potential has been introduced. A novel computational approach by incorporating the chemical potential with the protein binding probability is developed.

Materials and Methods

We have incorporated Bayesian model regularization with biophysical modeling of protein-DNA interactions, and of genome-wide nucleosome positioning to study protein-DNA interactions, using a high-throughput dataset. The newly developed method (BayesPI) includes the estimation of a transcription factor (TF) binding energy matrices, the computation of binding affinity of a TF target site and the corresponding chemical potential.

Results

The method was successfully tested on synthetic ChIP-chip datasets, real yeast ChIP-chip experiments. Subsequently, it was used to estimate condition-specific and species-specific protein-DNA interaction for several yeast TFs. The results revealed that the modification of the protein binding parameters and the variation of the individual nucleotide affinity in either recognition or flanking sequences occurred under different stresses and in different species.

Discussion

Our findings suggest that the modifications of protein binding parameters may be adaptive and play roles in the formation of the environment-specific binding patterns of yeast TFs and in the divergence of TF binding sites across the related yeast species.

URL

http://folk.uio.no/junbaiw/bayesPI/

Presenting Author

Junbai Wang (junbai.wang@rr-research.no)
Department of Pathology, Norwegian Radium Hospital, Oslo University Hospital

Author Affiliations

1. Department of Pathology 2. Department of Cell Biology, Institute for Cancer Research The Norwegian Radium Hospital, Oslo University Hospital Montebello 0310 Oslo, Norway

Acknowledgements

The Norwegian Cancer Society

E-12. Use of structural DNA properties for the prediction of regulator binding sites with conditional random fields

Meysman P (1,), Engelen K (1), Laukens K (2), Dang TH (2), Marchal K (1)*

Molecular recognition of genomic target sites by regulator proteins is a vital process in the transcription regulation of genes in living cells. The types of physical interactions that contribute to the recognition of binding sites by a protein can roughly be divided into those enabling direct read-out and those that allow for indirect read-out. The former comprises base-specific recognition, such as stabilizing hydrogen bonds between regulator amino acids and a set of conserved bases in the genomic DNA sequence, while in the case of the latter variations within the DNA structure will be used

Materials and Methods

The structural DNA properties of the target sites, needed to construct the model, are derived from their nucleotide sequence using a number of higher-order value look-up functions, so-called structural scales, which are based on experi-mental data (e.g. X-ray crystallography of various DNA molecules). These structural properties were used as input data to train a model representing the common structural features shared by all known binding sites of a specified regulator. This was done using conditional random fields (CRF), a discriminative machine learning method. Two novel extension algorith

Results

The performance of the presented methodology was demonstrated on data sets for 27 different Escherichia coli transcription factors and showed improved performance when compared to previously developed methods. Further a set of novel target site predictions, resulting from use of the trained models in a genome wide screening of E. coli, were validated using a microarray compendium of ca. 1500 arrays, as well as comparison with current literature.

Discussion

Using the presented framework, we created regulator target site models that showed a higher accuracy than similar methods and were able to find coded structural patterns that had been identified during crystallographic analysis of the protein-DNA complexes. Further we identified several new target sites using the structure-based models that were not previously uncovered using traditional sequence-based methods.

Presenting Author

Pieter Meysman (pieter.meysman@biw.kuleuven.be)
K.U.Leuven

Author Affiliations

(1) Department of Microbial and Molecular systems, K.U.Leuven, Kasteelpark Arenberg 20, B-3001 Leuven Heverlee, Belgium. (2) Intelligent Systems Laboratory, Department of Mathematics and Computer Science, Middelheimlaan 1, B-2020 Antwerpen, Belgium.

E-13. Gene expression analysis in Scots pine in response to red and far-red light

Ranade SS (*), Abrahamsson S, Niemi J, García-Gil MR

Light is vital for plant growth, morphogenesis and metabolism. Genome expression pattern under red light and far-red light is extensively studied in angiosperms but little is known about the gene regulation in this aspect with reference to Scots Pine

Materials and Methods

Seeds from northern Sweden were grown under continuous red and far-red light under two separate experiments. Microarray was performed with Pine cDNA microarrays. Microarray data was normalised by LOWESS normalisation method and by computing the log ratios of the intensity measurements. Statistical analysis was carried by calculating the t-test statistic and the corresponding p-values from the M-values for each gene in the dye swap experiment. The p-values were adjusted by FDR method using the R-statistical package. Computational annotation of the genes was done using Blast2GO

Results

405 genes were found to be up-regulated under red light treatment and down-regulated under far-red light; and 239 genes were identified which were down-regulated under red light and up-regulated under far-red light treatment. Microarray analysis thus reveals that there exist genes, which are differentially expressed and regulated under the red and far-red light treatments in the northern population of Scots pine in Sweden

Discussion

Phytochromes and cryptochromes are the well studied photoreceptors in plants that perceive red/far-red and blue light respectively. Phytochromes control the elongation of the hypocotyls and root; while cryptochromes are involved in the circadian rhythm of plants. In the current investigation, cryptochromes are found to be down regulated in red light treatment and upregulated under far-red light. Significant regulation of phytochromes was not detected in the study. This suggests that phytochrome regulation does not happen at the transcription level but rather at the more downstream stages

Presenting Author

Sonali S. Ranade (Sonali.Ranade@genfys.slu.se)
Department of Forest Genetics and Plant Physiology, Umea Plant Science Centre, Swedish University of Agricultural Sciences

Author Affiliations

Department of Forest Genetics and Plant Physiology, Umea Plant Science Centre, Swedish University of Agricultural Sciences (SLU), SE-901 83 Umeå, Sweden

E-14. Transcription regulatory regions and motifs (TRAM) database: the genomic data layer of the Nencki regulatory genomics portal

Dabrowski M (*), Lenart J, Mieczkowski J, Kaminska B

The vertebrate putative regulatory regions available in Ensembl are of several different kinds. Their relative (or joint) usefulness in understanding gene co-regulation merits further studies. The relative performance of existing libraries of transcription factor binding motifs, both public (JASPAR) and commercial (TRANSFAC, Genomatix) on the same genome wide-set of putative regulatory sequences is also of interest.

Materials and Methods

We designed and implemented a database schema unifying representation of the data from Ensembl funcgen, compara and core to the application layer above it. Several types of putative regulatory regions (VISTA Enhancers, cisRED promoters, AVID-VISTA or BlastZNet alignments) will be separately scored with TRANSFAC, Genomatix and JASPAR, using their local installations. The results will constitute the data layer used by the future application layer - together with user-supplied expression data, for analysis of gene co-regulation.

Results

We have completed AVID-VISTA non-coding alignments in the genomic sequence flanks of human-rat orthologous genes from Ensembl v.57, scored with the Genomatix motif library v. 8.2. This section of TRAM contains 43583 conserved non-coding regions for 9217 genes, with nearly 2 mln instances of 743 motifs from 180 families in either species. Scoring of all types of putative regulatory regions with TRANSFAC and Genomatix should soon be finished. We are also working on automation of TRAM updating.

Discussion

The unifying database schema permits comparison of different types of predicted regulatory regions with the experimentally verified VISTA Enhancer set and genome-wide chromatin modification data. The results obtained with the three motif libraries can be compared with the ChIP data for individual transcription factors. With the future application layer, it will be possible to compare the performance of different types of putative regulatory regions, and of the three motif libraries, in prediction of gene expression from the genomic sequence.

Presenting Author

Michal Dabrowski (m.dabrowski@nencki.gov.pl)
Nencki Institute

Author Affiliations

Nencki Institute, Laboratory of Transcription Regulation, Poland

Acknowledgements

This project is supported by POIG 02.03.00-00-003/09.

E-15. The microRNA body map: dissecting microRNA function through integrative genomics

Mestdagh P (1,#), Lefever S (1,#,), Pattyn F (1,2), Ridzon D (3), Fredlund (1), Fieuw A (1), Vermeulen J (1), De Paepe A (1), Wong L (3), Speleman F (1), Chen C (3), Vandesompele J (1)*

MicroRNAs (miRNAs) are small non-coding RNA molecules that function as indispensible regulators of an increasing number of cellular processes. The role of miRNAs is depending on its spatiotemporal expression pattern and the targeted genes. While miRNA expression profiles have been established for various normal and diseased tissues, our understanding of miRNA function remains limited. Here, we report a unique repository of published and unpublished miRNA expression data from more than 550 samples, which can serve as a resource for predicting miRNA function through integrative transcriptomics.

Materials and Methods

Unlike classical approaches, which rely solely on in silico miRNA target prediction to infer function, we combined genome wide mRNA and miRNA expression data in a correlative analysis and used gene set enrichment analysis (GSEA) to identify significant miRNA – gene set associations. Besides miRNA function and specific expression, the miRNA body map also enables the identification of stably expressed reference miRNAs in specified tissues, differential miRNAs between defined sample groups and candidate therapeutic miRNAs.

Results

To assess the validity of the inferred functions, GSEA-results were analyzed for tissue and sample subset specific miRNAs. These results show a striking correlation between the miRNA expression in these samples and the gene set annotation to which they are linked.

Discussion

The miRNA body map tailors each of the above described analyses according to user-defined settings in the available datasets. Our results demonstrate the value of integrated analysis and highlight the potential of the miRNA body map project as a community resource. To promote expansion of the current database, additional datasets can be uploaded.

Presenting Author

Steve Lefever (steve.lefever@ugent.be)
Center for Medical Genetics Ghent

Author Affiliations

(1) Center for Medical Genetics, Ghent University Hospital, Ghent, Belgium. (2) German Cancer Research Center (DKFZ), Heidelberg, Germany. (3) Life Technologies, Foster City, California, USA. (#) Equally contributing authors

E-16. D-Light: transcription factor binding site management and visualization

Laimer J (1,), Zuzan C (2), Ehrenberger T (1), Freudenberger M (1), Gschwandtner S (1), Lebherz C (1), Lirk G (1), Lackner P (2)*

Different databases like JASPAR, TRANSFAC or PAZAR collect experimental data about transcription factor binding sites (TFBSs) and provide the data as raw DNA sequences or positional frequency matrices (PFMs) for hundreds of different transcription factors (TFs). A number of computational methods have been developed to use these data for the prediction of potential binding sites on promoter sequences. This results in a large number of TF annotation data even for a single organism which need to be stored, queried and combined an efficient manner.

Materials and Methods

The server imports data from JASPAR (PFMs), UCSC (promoter sequences) and NCBI (homology information). The corresponding Python scripts can easily be adopted for other data sources. The server requires a Unix like OS. The client is written in Java for platform independency. The visualization is based on the GenoViz SDK.

Results

We present a client-server system for compiling, managing, querying and displaying annotation data for transcription factor binding sites, called 'D-Light'. The system provides simple access to the annotation data. Combinatorial queries within or between genomes allow for reasonable localization of target genes controlled by certain TFs. Moreover, custom data such as PFMs and promoter sequences can be added by the user. Finally, a user management enables privacy.

Discussion

We implemented a new client-server system for working with transcription factor binding sites in a large scale. Users can easily integrate custom data to be used in simple but also complex cross species combinatorial queries. D-light is provided as web service on http://biwww.che.sbg.ac.at/dlight. The server software is also available for local installation from the same web page.

URL

http://biwww.che.sbg.ac.at/dlight

Presenting Author

Josef Laimer (josef@laimer.cc)
Upper Austria University of Applied Sciences - School of Informatics, Communications and Media

Author Affiliations

(1) Upper Austria University of Applied Sciences - School of Informatics, Communications and Media (2) University of Salzburg, Department of Molecular Biology

Acknowledgements

Austrian Study Grant Authority

E-17. Expression and chromosomal clustering of tissue restricted antigens

Dinkelacker M(1,),Pinto S(2),Derbinski J(2),Eils R(1,3),Kyewski B(2),Brors B(1)*

Eliciting the molecular mechanism of the negative selection of T cells in the thymus gives raise to help understand and hopefully cure autoimmune diseases in the future. Tissue restricted antigens are promiscously expressed by medullary thymic epithelial cells (mTECs) which mirror virtually all tissues of the body to the T cells so that immune system will protect the body from pathogens but at the same time tolerate the bodys own tissues. Chromosomal clustering of TRAs gives a good explanation, how this monitoring could be regulated on a cellular level.

Materials and Methods

TRAs where defined by a bioinformatical approach to detect tissue specifically expressed genes in gene expression data in the mouse, chromosomal clustering was tested with a sliding gene window of fixed size and a sliding 10 gene window method, proposed by Roy et al. 2002 and Gotter et al. 2004

Results

With our bioinformatical approach of defining TRAs we could detect 5800 (25%) of all genes on the microarray to be tissue specific, and show that tissue restricted antigens are significantly clustered on the chromosome compared to 1000 times randomly drawn gene lists of the same length. Comparing this data to gene expression data measured in medullary thymic epithelial cells (mTECS) in AIRE ko versus wt mice, which suffer from multiple autoimmune diseases, we could show that our theoretically definition performed well to detect TRAs in microarray data.

Discussion

Chromosomal clustering of TRAs in the thymus can be an explanation of how T cells are trained to tolerate the bodys own tissues and how one single cell type in the thymus is able to express all different tissue type specific genes at once. A detailed picture of single TRA clusters proofs that single clusters are intermingled with genes of all different tissue types. Further analysis will hopefully help to understand how central tolerance is established and autoimmune diseases can be prevented.

Presenting Author

Maria A. Dinkelacker (m.dinkelacker@dkfz.de)
German Cancer Research Center, Heidelberg, DKFZ

Author Affiliations

(1)Dept. Theoretical Bioinformatics,German Cancer Research Center(DKFZ),Heidelberg, Germany (2) Dept. Developmental Immunology,German Cancer Research Center(DKFZ) Heidelberg, Germany (3) Department for Bioinformatics and Functional Genomics, Institute for Pharmacy and Molecular Biotechnology, Bioquant, University of Heidelberg, Heidelberg, Germany

E-18. Community dynamics and metatranscriptome analysis of lactic acid bacteria ecosystems through functional gene microarray analysis

Weckx S (1,), Allemeersch J (2), Van der Meulen R (1), Vrancken G (1), Huys G (3), Vandamme P (3), Van Hummelen P (2), De Vuyst L (1)*

LAB are of industrial importance in the production of fermented foods, among which sourdough-derived products. LAB contribute considerably to important characteristics of these fermented foods, such as an extended shelf-life and microbial safety, improved texture, and enhanced organoleptic properties. Thanks to the genomic information on LAB that became available during the last years, transcriptome and by extension metatranscriptome studies are the exquisite research approaches to study LAB whole-ecosystem gene expression and community dynamics into more detail.

Materials and Methods

Microarray hybridizations were performed using time-related RNA samples, representing the metatranscriptome, of four 10-day sourdough fermentations, and an in-house developed functional gene LAB microarray, consisting of 2,269 oligonucleotides targeting 406 genes that play a key role in sugar and nitrogen metabolism, functional metabolite production, stress responses, and safety characteristics. For each gene, a net expression profile was calculated. Microarray hybridization data were used to determine LAB community dynamics. Results were compared with data of metabolite target analyses.

Results

Hybridization data analysis revealed that the sourdough fermentations could be considered as a microbial ecosystem that achieved stability after a transition phase of five days, with Lactobacillus plantarum, Lactobacillus fermentum, and Pediococcus pentosaceus as the ultimate dominating species. The net expression profiles showed that gene expression was characterized by the activation of different key metabolic pathways, the ability to use different energy sources, and the conversion of amino acids as contribution to redox equilibrium and flavor compound generation.

Discussion

Overall gene expression of four sourdough fermentations were successfully followed up using a functional gene LAB microarray. The strains present in the sourdoughs showed an efficient use of the available substrates, such as the consumption of other carbohydrates than glucose, either as energy source or as alternative external electron acceptor, thereby elaborating metabolites of interest from a technological point of view. In addition, the hybridization patterns showed that the microarray could be used as alternative tool to monitor population dynamics during sourdough fermentations.

Presenting Author

Stefan Weckx (stefan.weckx@vub.ac.be)
Vrije Universiteit Brussel - IMDO

Author Affiliations

(1) Research Group of Industrial Microbiology and Food Biotechnology (IMDO), Faculty of Sciences and Bio-Engineering Sciences, Vrije Universiteit Brussel, Belgium (2) Microarray Facility, Flanders Institute for Biotechnology (VIB), Belgium (3) Laboratory of Microbiology, Faculty of Sciences, Ghent University, Belgium

E-19. A functional gene microarray for lactic acid bacteria exceeding the species level: design, validation, and data analysis

Allemeersch J (1,), Van der Meulen R (2), Vrancken G (2), Huys G (3), Vandamme P (3), Van Hummelen P (1), De Vuyst L (2), Weckx S(2)*

Functional gene microarrays allow to monitor the expression of a selected set of genes within a microbial ecosystem as a whole. To investigate whole-ecosystem gene expression during food fermentations, a functional gene microarray for LAB was developed targeting 406 genes across 50 species that play a key role in carbohydrate metabolism, in the production of bacteriocins, exopolysaccharides, and flavour compounds, in probiotic and biosafety characteristics, and in stress response. Also, genes linked to negative traits such as antibiotic resistance and virulence were represented.

Materials and Methods

As LAB ecosystems can contain a variety of species, the microarray design focused on functional properties, beyond the species level. An algorithm was implemented to design gene-specific 30-mer oligonucleotides that preferably could hybridize multiple LAB species, to allow for controlled cross-hybridization. Validation hybridizations were performed using RNA and DNA of 18 LAB strains. Also, a new algorithm was developed to analyze the hybridization data by calculating a net expression profile for each gene represented on the microarray, thereby exceeding the species level.

Results

Applying the oligonucleotide design algorithm on manually curated public LAB sequence data resulted in 4,851 oligonucleotides. A subset of 2,269 oligonucleotides was selected for synthesis and hence production of the microarray based on the LAB microbiota prevailing during sourdough fermentations. This subset represented 406 key genes in a total of 46 LAB and four non-LAB species. The validation hybridizations using DNA and RNA from 18 strains of twelve LAB species, representing 86.0% of all oligonucleotides, showed wide ranges of intensity and high reproducibility for technical replicates.

Discussion

A species-independent functional gene LAB microarray was developed to study gene expression of LAB in fermented food ecosystems. An algorithm was implemented to design 30-mer oligonucleotides with controlled cross-hybridization capabilities between species for a given gene. Of the 2,269 oligonucleotides on the microarray, 15.5% could cross-hybridize multiple species. Validation hybridizations proved that 30-mer oligonucleotides showed good sensitivity and reproducibility. The net expression profiles allowed to interpret whole-ecosystem gene expression, exceeding the individual species level.

Presenting Author

Joke Allemeersch (joke.allemeersch@vib.be)
VIB-MAF

Author Affiliations

(1) Microarray Facility, Flanders Institute for Biotechnology (VIB), Belgium (2) Research Group of Industrial Microbiology and Food Biotechnology (IMDO), Faculty of Sciences and Bio-Engineering Sciences, Vrije Universiteit Brussel, Belgium (3) Laboratory of Microbiology, Faculty of Sciences, Ghent University, Belgium

E-20. Combinatorial analysis of transcription regulation of response to oxidative stresses in E. coli

Thiele S (1), Klie S (3), Jozefczuk S (2), Selbig J (2, 3), Blachon S (3,)*

Transcriptional regulation is a crucial component of bacterial stress response. The existence of operons and global transcriptional regulators allows E. coli to quickly adjust to changing environment. However the great variability of environmental signals requires the ability to combine activity of different transcriptional regulators to fine tune the response. By integrating (anti)correlated gene responses with knowledge on transcription factor regulation we aim at detecting behaviors that cannot be explained with current knowledge and can result from more complex yet unknown regulations.

Materials and Methods

Using microarrays, the time response to various stresses was measured at the transcript level. Correlations between genes that vary under stress were computed. When two genes are (anti)correlated, at least one common transcription factor should be responsible for the coordinated response. Conversely, if two genes have no common transcription factor, they should not be (anti)correlated. We propose an original approach based on the Answer Set Programming (ASP) framework. ASP enables to model the problem, detect conflicts and propose sets of solutions that resolve the conflicts.

Results

Focusing on oxidative stress response, 264 genes vary significantly and have at least one known transcription factor regulating their activities. Overall, 83 transcription factors regulate these genes. Applying the ASP framework, conflicts were detected in 4471 gene pairs. Using a feature of ASP, the cardinality constraint, it is possible to compute the minimal number of repairs. Here, adding 3 unknown transcription factors is enough to repair all conflicts. These unknown transcription factors can be identified as real transcription factors for which the targets remain to be discovered.

Discussion

Exploring all possible combinations of repairs can be done only at a high computational cost. We are currently implementing new rules to refine the search space. These rules are based on the problem properties and on biological knowledge (common region in promoter sequences). So far, the approach was applied only to data produced on oxidative stress. Other stresses (heat shocks, ...) will be studied in the short run.

Presenting Author

Sylvain Blachon (sylvain.blachon@gmail.com)
Department of Bioinformatics, University of Potsdam, Potsdam, Germany

Author Affiliations

(1) Institute for Computer Science, University of Potsdam, Potsdam, Germany (2) Max Planck Institut for Molecular Plant Physiology, Potsdam, Germany (3) Institute for Biochemistry and Biology, University of Potsdam, Potsdam, Germany

Acknowledgements

GoFORSYS project

E-21. Time delays in gene activation: sensitivity to external noise

Albert J (1,), Rooman M (1)*

Many biological processes occur through a controlled sequence of events – a cascade – in which the activation of one gene by another is delayed. Understanding the mechanisms that cause delays is of great importance in, among other fields, developmental biology where the ability of cells to turn on certain genes at specific times plays a crucial role. How the activation of genes can be delayed with reliable accuracy even in the presence of external perturbations needs to be addressed to ensure some biological relevance.

Materials and Methods

We have modeled the dynamics of a genetic switch by a set of two coupled differential equations, using mathematica. The parameters of the system were perturbed, one at a time, by Gaussian noise occurring at time intervals whose probability followed a Poisson distribution. The statistics of the switch were carried out on a basis of a pool of fifty runs for several combinations of average amplitude and time intervals of the noise. The performance of the switch was judged on the basis of how close were its average delays to the exact delays (without noise) and their standard deviations.

Results

We have found that the robustness of the delays against random parameter perturbations is determined primely by the ratio of protein to mRNA degradation rates. The larger this ratio, the more sensitive are the delays to parameter fluctuations. The performance of the switch was observed to be dependent on the particular parameter that was subjected to noise. It turns out that the optimal way to generate time-insensitive delays is through the binding of an activating transcription factor, which does not interact with the protein synthesized from the gene under consideration.

Discussion

Considering the range of performances given by the genetic switch we found that only some combinations of parameters allow the activation of the switch to be delayed with reliable accuracy. The observations that the robustness of the delays against noise depends strongly on the degradation rates of protein and mRNA, and that these rates can differ by orders of magnitude from eukaryotes to prokaryotes, suggest that the switching delays caused by the proposed mechanism may be unique to one or the other type of cell.

Presenting Author

Jaroslav Albert (jalbert@ulb.ac.be)
Université Libre de Bruxelles

Author Affiliations

Biosystems, Biomodeling and Bioprocesses, Université Libre de Bruxelles

E-22. Discovering regulatory mechanisms that govern zygotic genome activation in Drosophila

Darbo E (1,), Thieffry D (1), Lecuit T (2), van Helden J (3,1)*

In Drosophila embryos, no transcription occurs during the first 7 cell cycles after fertilization. The control of the early steps of development is ensured by maternal mRNAs loaded in the egg during oogenesis. Zygotic genome activation (ZGA) occurs in two waves: the first wave involves about 60 genes, whose transcription is activated during mitotic cycle 8; a second wave involving more than 300 genes occurs at the 14th mitotic cycle. We aim at characterizing the mechanisms involved in the control of transcriptional activation during ZGA.

Materials and Methods

We integrated transcriptome profiles for 3 independent studies focused on (i) gene transcription during early embryogenesis [Pilot et al. (2006) Development 133:711]; (ii) respective contributions of maternal and embryonic genes [De Renzis et al.(2007) PLoS Biol 5:e117]; (iii) impact of nucleo/cytoplasmic ratio on transcription [Lu et al. (2009) Development 136: 2101]. We defined clusters of genes showing significant changes in expressions at crucial stages, predicted their putative cis-regulatory motifs, and detected functional enrichment using Gene Ontology (GO) annotations.

Results

We analyzed 73 clusters of co-expression and discovered highly significant motifs. Some of the discovered motifs correspond to the consensus of known transcription factors (e.g. CAGGTAC is the consensus of Zelda, involved in the first wave of ZGA). For some clusters, the transcription factors predicted by motif discovery (Zelda, Dref) were consistent with the functional enrichment in GO classes related to embryonic development. We also found novel motifs (e.g., CAGATACA) which do not correspond to any annotated factor.

Discussion

We developed a pipe-line combining the analysis of transcriptome data, regulatory motifs and functional enrichment. The finding of known motifs such as CAGGTAG shows that the approach is able to return relevant results, and increases our interest for the novel motifs discovered. Further analyses of combination of motifs (motif co-occurrences, detection of cis-regulatory modules) and multi-genome analyses (conservation and divergence of putative binding sites) will contribute to select the most promising motifs as candidate for experimental validation.

Presenting Author

Elodie E Darbo (darbo@tagc.univ-mrs.fr)
TAGC INSERM U928

Author Affiliations

1.Technological Advances for Genomics and Clinics (TAGC), INSERM U928 & Université de la Méditerranée (Aix-Marseille II). Campus de Luminy, Case 928, F - 13288 Marseille, France. darbo@tagc.univ-mrs.fr, thieffry@tagc.univ-mrs.fr 2.Laboratoire de Génétique et de Physiologie du Développement, UMR 6545 CNRS-Université de la Méditerranée, Institut de Biologie du Développement de Marseille (IBDM), Campus de Luminy, case 907, 13288 Marseille CEDEX 09, France. lecuit@ibdm.univ-mrs.fr 3.Laboratoire de Bioinformatique des Génomes et des Réseaux (BiGRe). Université Libre de Bruxelles, Campus Plaine, CP 263. Bld du Triomphe, B-1050 Bruxelles, Belgium. Email: Jacques.van.Helden@ulb.ac.be

Acknowledgements

This work is supported by the ANR Program NeMo. The collaboration between BiGRe and TAGC laboratories is supported by the Interuniversity Attraction Poles, funded by the Belgian Federal Science Policy Office, project P6/25 (BioMaGNet).

E-23. Detection of extended UTRs with 3' expression microarrays

Thorrez L (1,), Tranchevent L-C (1), Chang H-J (1), Moreau Y (1), Schuit F (1)*

The 3' untranslated regions (UTRs) of transcripts are not well characterized for many genes and often extend beyond the annotated regions. Since Affymetrix 3' expression arrays were designed based on expressed sequence tags, many probesets map to intergenic regions downstream of genes. We used expression information from these probesets to predict transcript extension beyond currently known boundaries.

Materials and Methods

The expression dataset consisting of 70 microarrays covering 22 different murine tissues with 3-5 replicates per tissue was used as starting data. Data are accessible through the GEO database, with accession number GSE9954 [32]. Analysis was carried out with scripts in R, Perl and Java. Confimation of predicitions was performed by PCR.

Results

Based on our dataset encompassing expression in 22 different murine tissues, we identified 845 genes with predicted 3'UTR extensions. These extensions have a similar conservation as known 3'UTRs, which is distinctly higher than intergenic regions. We verified 8 of the predictions by PCR and found all of the predicted regions to be expressed. The method can be extended to other 3' expression microarray platforms as we demonstrate with human data. Additional confirming evidence was obtained from public paired end read data.

Discussion

We show that many genes have 3'UTR regions extending beyond currently known gene regions and provide a method to identify such regions based on microarray expression data. Since 3' UTR contain microRNA binding sites and other stability determining regions, identification of the full length 3' UTR is important to elucidate posttranscriptional regulation.

Presenting Author

Lieven Thorrez (lieven.thorrez@esat.kuleuven.be)
KULeuven

Author Affiliations

(1) SymBioSys - KUL Center for computational systems biology

Acknowledgements

This research was supported by research grants from the Fonds voor Wetenschappelijk Onderzoek Vlaanderen (FWO grant G.0733.09N), Katholieke Universiteit Leuven (GOA/2004/11 to the FS laboratory, GOA AMBioRICS (2004-2009), GOA 3M080124 and SymBioSys: CoE EF/05/007) and the Belgian Federal Science Policy Office (IUAP P6/40 and P6/25: BioMaGNet). L. Tranchevent is funded by EU-RTD: ERNSI: FP6-MC-EST Bioptrain.

E-24. A dynamic regulatory network model reveals hard-wired heterogeneity in blood stem cells

Bonzanni N (1,), Garg A (2), Foster SD (3), Wilson NK (3), Kinston S (3), Miranda-Saavedra D (3), Heringa J (1), Feenstra A (1), Xenarios I (2), Göttgens B (3)*

Combinatorial interactions of transcription factors with cis-regulatory elements control the dynamic progression through successive cellular states and thus underpin all metazoan development. The construction of regulatory network models based on the functionality of cis- regulatory elements therefore has the potential to generate fundamental insights into cellular fate and differentiation. Haematopoiesis has long served as a model system to study mammalian differentiation, yet modelling based on experimentally informed interactions has so far been restricted to pairs of interacting factors.

Materials and Methods

Here we have generated a network model based on detailed cis-regulatory functional data connecting 11 haematopoietic stem cell (HSC) regulators. We used Boolean logic functions to model the interactions between the cis-regulatory elements. F0 transgenic mouse embryos were generated and analyzed to test the novel hypotheses.

Results

Dynamic analysis of our model predicts that HSCs display heterogeneous expression patterns and possess many intermediate states that appear to act as 'stepping stones' for the HSC to achieve a final differentiated state. By focussing on intermediate states occurring during erythrocyte differentiation, we predicted a novel negative regulation of Fli1 by Gata1 which we confirmed experimentally thus validating our model.

Discussion

In conclusion, we present the most advanced mammalian regulatory network model based on experimentally validated cis-regulatory interactions to date. This model has allowed us to make novel, experimentally testable hypotheses about transcriptional mechanisms that control differentiation of mammalian stem cells.

Presenting Author

Nicola Bonzanni (bonzanni@few.vu.nl)
Vrije Universiteit Amsterdam

Author Affiliations

1 IBIVU/Bioinformatics, Free University Amsterdam, De Boelelaan 1083A - 1081 HV, Amsterdam, The Netherlands 2 Swiss Institute of Bioinformatics, Quartier Sorge - Batiment Genopode, CH-1015 Lausanne, Switzerland 3 Cambridge Institute for Medical Research, University of Cambridge, Hills Road, Cambridge CB2 0XY, UK

Acknowledgements

This study was funded by the European Commission FP6 project ENFIN (Experimental Network for Functional INtegration). Computations were performed at the Vital-IT Center of the Swiss Institute of Bioinformatics. Experimental work was supported by Leukaemia & Lymphoma Research (UK), the Leukemia & Lymphoma Fund (USA) and the Medical Research Council (UK).

E-25. Alternative splicing as a marker of disease

Muro EM ( ), Andrade-Navarro MA*

The goal of this work is to identify alternative splicing variants depending on tissue and age, classifying them according to a ranked list of candidate genes for neurodegenerative diseases. In addition, we will investigate possible regulatory mechanisms for splice variation based on antisense expression. Here we present our latest advances in these subjects ending with the analysis of alternative splicing related to disease.

Materials and Methods

For detecting alternative splicing variants we use different approaches: for the 3' UTR alternative variants we profit from our own experience with "Transcriptome Sailor" (see Muro E.M. et al. 2008. Identification of gene 3' ends by automated EST cluster analysis. Proc. Natl. Acad. Sci. 150:20286-20290). Data from the Alternative Splicing and Transcript Diversity database 1.1 is used in combination with the corresponding Ensembl genome annotation. For analyzing tissue, age and some other factors associated to transcripts we use the Genebank annotations of the cDNA data (ESTS and mRNA) in combination with MeSH terms.

Results

In a first analysis we detected 3,875 genes (4,187 splice variants) related to the central nervous system using the different databases mentioned above. Independently we observed that some NATs (natural antisense transcripts) arise from pseudogenes, identifying 87 cases genome wide (human). In particular, we hypothesize that some of them might be regulating splicing, because they bind pre-mRNAs. For some exons that are subject of exon skipping events, we found cis-NATs expressed in the regulatory regions (cis-acting receptors) of the skipped exon, probably acting as splicing regulators.

Discussion

A large percentage of human diseases are caused by aberrant alternative splicing. In some other cases, alternative splicing isoforms are transcribed as a consequence of a disease but do not cause it. In the last case the alternative splicing variants would be markers of the disease. We are collecting splicing variants, classifying them depending on factors like tissue or age. We hope to obtain splicing variants that are unique in certain age range tissue and disease. In parallel we are working on the regulation of the splicing variants always with the aim of finding examples that could be related to disease.

URL

http://cbdm.mdc-berlin.de/

Presenting Author

Enrique M Muro (enrique.muro@mdc-berlin.de)

Computational Biology and Data Mining Group Max-Delbrueck-Centrum fuer Molekulare Medizin (MDC). Berlin. Germany.

Author Affiliations

Computational Biology and Data Mining Group Max-Delbrueck-Centrum fuer Molekulare Medizin (MDC). Berlin. Germany.

Acknowledgements

MSBN, NGFN.

E-26. TilSeg: an automated pipeline to process tiling array expression data

Barreau D (1,3,4), Gao Y (2,3,4), Jaffrezic F (2,3,4), Hugo K (1,2,3), Rogel-Gaillard C (2,3,4), Marthey S (1,2,3,)*

High density oligonucleotide tiling arrays represent a powerful approach to decipher gene expression in genomic regions that are partially or not annotated. However, data analysis remains challenging due to the lack of free or commercial softwares. Moreover, most of the existing methods focus on probe expression levels and do not take into account the information provided by contiguous probes. We thus developed a method to take advantage of this kind of array design and a tool to process automatically the data.

Materials and Methods

TilSeg mainly uses perl for managing files and for automatized treatment of data. This tool uses standard R packages to perform statistical analysis of the data, including normalization and segmentation methods.

Results

We developed a user friendly tool which can be used directly on a batch of quantification files. The analysis of tiling arrays data can be easily processed using the standard parameters defined in TilSeg, but if needed all the statistical parameters can be easily modified. In addition, output files can be produced after each step of the statistical pipeline and used directly as input files for other analysis tools. Finally the data fully processed through TilSeg can be formated and visualized directly in current genome browsers as Ensembl or UCSC.

Discussion

TilSeg can be used as an automated pipeline for basic statistical analysis of Tiling Array expression data. The new methods of analysis that are currently being developed will be implemented in TilSeg as new functionalities, making it a sustainable tool for the analysis of tiling arrays. Further developments should enlarge the scope of TilSeg, including for example automatic annotation of segments of interest and the analysis of tiling arrays for CNV characterization.

Presenting Author

Sylvain Marthey (sylvain.marthey@jouy.inra.fr)
CRB GADIE / GABI / INRA

Author Affiliations

1 INRA, CRB GADIE, UMR de Génétique Animale et Biologie Intégrative, Jouy-en-Josas, France 2 INRA, UMR de Génétique Animale et Biologie Intégrative, Jouy-en-Josas, France 3 CEA, DSV, iRCM, Laboratoire de Radiobiologie et Etude du Génome, Jouy-en-Josas, France 4 AgroParisTech, UMR de Génétique Animale et Biologie Intégrative, Jouy-en-Josas, France

E-27. Evaluation of computational miRNA target predictions in human

*Gevaert O**

MicroRNAs (miRNAs) are short RNAs that regulate expression through binding to the 3’UTR of mRNAs. An important shortcoming in current miRNA research is the lack of experimentally verified mRNA targets. This initiated a surge in the development of computational methods attempting to predict miRNA target sites.

Materials and Methods

Here, we compare the target predictions of seven often used target prediction tools for human miRNAs: microRNA.org, MicroCosm, PITA, TargetScan, PicTar, MirZ and MicroT.

Results

Our results showed that most tools provide target predictions for the majority of currently known miRNAs but large differences where observed in the number of targets predicted on average. Comparison with experimentally verified targets showed that approximately half of the verified interactions are not predicted by most tools. Finally, we investigated the coordinated downregulation and showed that MirZ and MicroT captured more significant miRNA mediated down-regulation compared to the other tools in a wide range of tissues.

Discussion

We conclude that many of the tools predict less than half of the experimental observations. In addition, many tools predict different targets for the same miRNA. The availability of more experimentally verified miRNA-mRNA interactions could boost the development of target prediction models significantly and allow a better characterization of the miRNA-mRNA-interactome.

URL

http://www.esat.kuleuven.be/~bioiuser/microRNA/

Presenting Author

Olivier Gevaert (olivier.gevaert@esat.kuleuven.be)
Stanford/KU Leuven

Author Affiliations

Stanford University School of Medicine Department of Radiology/Katholieke Universiteit Leuven, Department of Electrical Engineering

E-28. Model-Based Multifactor Dimensionality Reduction to detect epistasis for quantitative traits in the presence of error-free and noisy data

Mahachie John JM (1,2,), Van Lishout F (1,2) , Van Steen K (1,2)*

Cattaert et al. (2010) have shown the excellent power of MB-MDR over MDR (Ritchie et al. 2003) in identifying epistasis effects on dichotomous traits in the presence of noisy data. Although MB-MDR has been successfully applied to quantitative traits as well by Mahachie John et al (2009), the power of MB-MDR in identifying significant epistasis effects on a quantitative trait in the presence of noisy data has never been explored.

Materials and Methods

We use a simulation study to evaluate the power and type 1 error rates of MB-MDR for quantitative traits in the presence of noisy data. Considered sources of error are genotyping errors, missing genotypes, phenotypic mixtures and genetic heterogeneity. The MB-MDR strategy to tackle the dimensionality problem involves reducing the dimension to one by pooling multi-locus genotypes into three groups of risk : High risk (H), Low risk (L) or No Evidence for risk (O). An association test is then performed with the new predictor variable X in {H,L,O} on the outcome variable.

Results

The obtained results show that MB-MDR has excellent power to identify epistasis effects in the presence of missing genotypes and low percentages of genotyping errors. Lowest power performances were observed in the presence of phenotypic mixtures and genetic heterogeneity. For null data with phenotypes generated completely at random, the type I error percentages are very close to the nominal level of 5%.

Discussion

For some time to come, phenotypic mixtures and genetic heterogeneity seem to remain challenging for epistasis screening methods. This was also observed by Cattaert et al. (2010) for dichotomous traits. Despite the flexibility offered by the MB-MDR framework, more work is needed to improve the performance under phenotypic mixtures and complex genetic heterogeneity patterns

Presenting Author

Jestinah M Mahachie John (jmahachie@ulg.ac.be)
University of Liege- Institute Montefiore

Author Affiliations

(1) Montefiore Institute, University of Liege, Belgium (2) GIGA-Research, University of Liege, Belgium

Acknowledgements

J.M. Mahachie John is a doctoral student funded by the Belgian Network BioMAGNet (Bioinformatics and Modeling: from Genomes to Networks), within the Interuniversity Attraction Poles Program (Phase VI/4), initiated by the Belgian State, Science Policy Office. All authors acknowledge research opportunities offered by the Belgian Network BioMAGNetand partial support by the IST Program of the European Community, under the PASCAL2 Network of Excellence (Pattern Analysis, Statistical Modeling and Computational Learning), IST-2007-216886.

E-29. Time coherent three-dimensional clustering: unraveling local transcriptional patterns in multiple gene expression time series

Gonçalves JP (1,2,3,), Moreau Y (3), Madeira SC (1,2)*

Analyzing temporal expression programs is essential to the understanding of complex biological processes. In particular, interesting local patterns revealing coherent behavior of genes in the time frame of a biological event can be sought using biclustering. However, as the amount of available data increases, research studies have been progressing towards comparative analysis of multiple time series expression matrices (TSEMs). Few methods have been proposed to address this problem, the majority relying on pseudo simultaneous three-dimensional approaches where one dimension is never clustered.

Materials and Methods

Given a set of discretized TSEMs, all maximal triclusters of genes exhibiting coherent expression patterns in specific time frames and TSEMs are identified as follows: 1) Alphabet transformation to ensure time coherence; 2) Expression pattern matching using a generalized suffix tree of the transformed patterns to identify nodes containing triclusters; 3) Frequent closed itemset mining to reveal the maximal subsets of genes and matrices underlying on such nodes; 4) Integration of pattern matching and itemset mining results into triclusters; Removal of non-pattern-maximal triclusters.

Results

Preliminary results analyze a dataset consisting of gene expression measurements obtained for 2920 genes and 9 time points in a 168 hour period for 14 multiple sclerosis patients after injection of IFN-beta using GeneFilters GF211 DNA arrays. Expression levels were normalized by gene to mean 0 and standard deviation 1 and then discretized using a technique expressing the variations between time points. Triclusters were subsequently identified using the time coherent three-dimensional clustering algorithm.

Discussion

Preliminary studies show promising results. Interesting expression patterns shared by sets of patients could be identified unraveling putative gene modules involved in patients' response to the treatment with IFN-beta.

Presenting Author

Joana P. Gonçalves (jpg@kdbio.inesc-id.pt)
KDBIO / INESC-ID

Author Affiliations

(1) KDBIO / INESC-ID, Rua Alves Redol 9, 1000-029 Lisboa, Portugal (2) IST / Technical University of Lisbon, 1049-001 Lisboa, Portugal (3) BIOI / ESAT-SCD, Katholieke Universiteit Leuven, Kasteelpark Arenberg 10, B-3001 Leuven, Belgium

Acknowledgements

This work was partially supported by FCT (INESC-ID multiannual funding) through the PIDDAC Program funds and project ARN - Algorithms for the Identification of Genetic Regulatory Networks, PTDC/EIA/67722/2006. The work of JPG was supported by FCT doctoral grant SFRH/BD/36586/2007.

E-30. The role of incoherent microRNA-mediated feedforward loops in noise buffering

Osella M (1), Bosia C (1), Corà D (2), Caselle M (1,)*

MicroRNAs are endogenous non coding RNAs that play important gene regulatory roles in animals and plants by pairing to the messenger RNAs of protein-coding genes to direct their post-transcriptional repression. Transcriptional and miRNA regulations are interlinked in a complex network in which the microRNA-mediated feed forward loop is a remarkably overrepresented regulatory circuit. The aim of our work is to suggest a possible role of this network motif in controlling the fluctuations in the target protein levels

Materials and Methods

We used both analytical and numerical methods. In particular we described the circuit using a set of stochastic equations which we then solved using the generating function approach and then tested our results using Gillespie simulations.

Results

We show that with respect to the simple gene activation by a TF, the introduction of a miRNA-mediated repressing pathway can significantly dampen fluctuations in the target protein output and that this noise buffering function is a consequence of the peculiar topolgy of the FFL.

Discussion

Our model predicts that the optimal attenuation of fluctuations coincides with a modest repression of the target expression. This feature is coherent with a fine-tuning function and in agreement with experimental observations of the actual impact of a wide class of microRNAs on the protein output of their targets.

Presenting Author

Michele Caselle (caselle@to.infn.it)
Dep. of Theoretical Physics, Torino University

Author Affiliations

(1) Department of Theoretical Physics, University of Torino (2) Systems Biology Lab (IRCC) University of Torino

E-31. Dynamic transcriptomics of helper T cell differentiation

van den Ham HJ (1,2,), de Waal L (2), Zaaraoui F (2), Bijl M(2), van IJcken WF(3), Osterhaus ADME (2), de Boer RJ (1), Andeweg AC (2)*

Helper T cells can adopt a number of different phenotypes, including Th1, Th2, Th17, Th9, and Treg. We aim to study the gene interacting network controlling the induction and development of these Th phenotypes in time. Because the majority of genes change expression upon activation of T cells, it is difficult to find Th phenotype associated genes. The recently developed polar score method enables us to fish for differential genes in a large set of genes that change expression. Using this dynamic transcriptomics approach, we identify several candidate genes related to Th1/Th2 differentiation.

Materials and Methods

We performed microarray assisted mRNA profiling on antigen specifically stimulated, TCR transgenic murine splenocytes that were cultured in the presence of polarising cytokines. Transcriptome snapshots of Th cells differentiating into Th1 and Th2 phenotypes were obtained.

Results

Principle component analysis shows that: 1) activation, 2) time since activation, and 3) Th skewing are the largest sources of variance in our profiling experiments. Multiple immune regulatory pathways are differentially expressed, including pathways specific to Th differentiation. The dynamic transcriptomics approach, applying the polar score data analysis method, enabled us to identify novel clusters of genes associated with Th1 and Th2 differentiation.

Discussion

Among the Th1 and Th2 specific genes that we identify are a number of key players previously associated with other T lymphocyte phenotypes or the maturation of other lymphoid cell types. We hypothesise that many of these genes play a universal role in cell differentiation, and are not necessarily restricted to Th cells.

URL

http://www.virgo.nl

Presenting Author

Henk-Jan van den Ham (h.j.vandenham@erasmusmc.nl)
ErasmusMC, Rotterdam, The Netherlands

Author Affiliations

1 Theoretical Biology & Bioinformatics, Utrecht University, The Netherlands, 2 Dep. of Virology, 3 Erasmus Center for Biomics, Erasmus MC, Rotterdam, The Netherlands * Presenting & corresponding author: h.j.vandenham@erasmusmc.nl

Acknowledgements

This project was funded by the VIRGO consortium, an innovative cluster approved by the Netherlands Genomics Initiative and partially funded by the Dutch Government (BSIK 03012).

E-32. Effect of size and heterogeneity of samples on biomarker discovery: synthetic and real data assessment

Di Camillo B (1), Martini M (1), Sanavia T(1), Jurman G (2), Sambo F (1), Barla A (3), Squillario M (3), Furlanello C (2),Toffolo G (1,), Cobelli C (1)*

The identification of robust lists of molecular biomarkers related to a disease is a fundamental step for early diagnosis and treatment. However, methodologies for the discovery of biomarkers using microarray data often provide results with limited overlap. These differences are imputable to 1) dataset size; 2) heterogeneity of the disease; 3) heterogeneity of experimental protocols and computational analysis. Addressing issues 1 and 2, we assessed both on simulated and clinical datasets, the consistency of candidate biomarkers provided by a number of different methods.

Materials and Methods

Three datasets of breast cancer patients ER+/- were used to assess list stability. To assess average precision/ recall on a number of studies, microarray data were simulated based on a gene network simulator, accounting for intrinsic variability of the population, simulated based on a model of evolution, and heterogeneity of the disease, simulated by modifying subjects' regulatory network. SAM and different methods for binary classification and feature ranking (SVM, I-Relief, Spectral Regression Discriminant Analysis), with/without external cross validation, were applied to data.

Results

Results on simulated data show that when some tens of subjects are available per group, performance in terms of features ranking and feature selection is good independently of the method. However, when available subjects are equal or lower than 15, the multivariate machine learning methods improve both the precision and the recall. The stability of the lists of features obtained using SAM both on simulated and real data, is poorer than the one given by the other methods. The use of external cross-validation loops improves results reproducibility rather than precision and recall.

Discussion

Altogether, these results show that although biomarkers are difficult to be univocally identified, biological signatures are consistent across different datasets when appropriate methods are used for the analysis. In particular, classification based approaches improve both precision and recall in feature selection, whereas using external cross-validation loops improves the reproducibility of biomarker lists. The in silico approach allowed us to outline advantages and drawbacks of different methods in a variety of experimental conditions. Application to real data confirms these results.

Presenting Author

Gianna Toffolo (dicamill@dei.unipd.it)
University of Padova

Author Affiliations

(1) Information Engineering Department, University of Padova, via Gradenigo 6A, I-35131 Padova, Italy. (2) Fondazione Bruno Kessler, via Sommarive 18, I-38123 Povo (Trento), Italy. (3) Department of Computer and Information Science, University of Genova, via Dodecaneso 35, I-16146 Genova, Italy.

Acknowledgements

FUNDINGS This work was supported by CARIPARO 2008/2010 "Systems biology approaches to infer gene regulation from gene and protein time series data"; ATENEO 2009/2010 "Methods for the integration of background knowledge in kernel-based learning algorithms for the robust identification of biomarkers in genomics"; RF Project 2007 Conv. 42. ISITAD. ACKNOWLEDGEMENTS We would like to thank Dr. Marco Falda and Roberto Visintainer for their help in running the simulations.

E-33. Analysis of DNA sequence dependent structural properties in prokaryotic genomes

Rangannan V (1,), Bansal M (1)*

Sequence dependent structural properties and their variation in genomic DNA are important in controlling several crucial processes such as transcription, replication, recombination and chromatin compaction. As there is rapid accumulation of genomic data, quantitative analysis of sequences motifs as well as structural properties, such as curvature, bendability and stability in the upstream region of genes, has become very important in order to automate the annotation pipeline for the newly sequenced genomes.

Materials and Methods

Genomic sequences flanking the Transcription Start Sites (TSSs) in E. coli, B. subtilis and M. tuberculosis have been taken from Ecocyc, DBTBS and MtbRegList databases respectively. The stability of a double stranded DNA molecule has been calculated using the free energy values of dinucleotides steps. The bendability profiles were calculated using two trinucleotide model values (DNase I sensitivity and Nucleosomal positioning preference). Curvature has been calculated using the dinucleotide structural parameters from crystal structure analysis and relative gel mobility data.

Results

Promoter regions have a preponderance of AT-rich tetranucleotides in E. coli and B. subtilis, not in high GC containing M. tuberculosis. % occurrence of TATA containing sequences is less in promoter regions as compared to other AT-containing tetramers. Average stability profile showed a low stability peak near TSS in all three cases. Bendability profile using both DNase I sensitivity and nucleosomal positioning preference models showed less bendable region around TSS, but not for M. tuberculosis and other GC rich genomes. High curvature in promoter regions is more pronounced in E. coli.

Discussion

The lower occurrence of AT-containing tetramers in the vicinity of high GC containing promoters, may account for the absence of low bendability peak in M. tuberculosis. Analysis of the structural features in the promoter sequences of these three systems showed that lower bendability and stability are seen in an equal number of promoters in E. coli and B. subtilis. Curvature is observed in a smaller number of the promoter sequences as compared to other two structural properties. Similar conclusions can be drawn from analysis of regions upstream of genes in all prokaryotic genomes.

URL

http://nucleix.mbu.iisc.ernet.in/prombase

Presenting Author

Vetriselvi Rangannan (vetri@mbu.iisc.ernet.in)
Molecular Biophysics Unit, Indian Institute of science, Bangalore, INDIA

Author Affiliations

(1) Molecular Biophysics Unit, Indian Institute of science, Bangalore-560012, INDIA.

Acknowledgements

This work was supported by Department of Biotechnology, INDIA.

E-34. Computational analysis of transcription factor binding co-localization

Martinez P (1,), Blanc E (2), Coolen A (1), Holzwarth J (3), Fraternali F (1)*

Transcription factors (TFs) are proteins that regulate gene expression, therefore cooperativity among them provides a complex mechanism to accurately control many biological processes. TFs display different modes of cooperation, either directly binding to DNA or through protein interactions with various partners and under different conditions. Because of this variability in cooperative modes and the growing amount of data on protein-protein interactions and regulatory networks, the analysis and the validation of TFs' cooperation remains a challenge.

Materials and Methods

Scanning human and mouse promoter regions for fixed nucleotide distances between pairs of DNA binding targets led to detection of clears outliers. These outliers repesented set of genes for whom co-regulation was evaluated using microarray data. In the last stage of the analysis, the Reactome database of known pathways was used to characterize the most interesting results.

Results

Evidence of transcription factor cooperation based on DNA sequence in promoter regions was found to bear consequences observed at the protein expression stage at a statistically significant level. Results shed light on the regulation of many known pathways and may be interesting hints for new functional pathways yet to be annotated.

Discussion

The results obtained in our analysis confirm the efficiency of our method for the detection of over-represented distances between pairs of DNA binding targets. Furthermore, the use of gene expression and known pathways data allowed us to suggest cooperation between transcription factors for regulation known and unknown pathways. Our results would still need experimental validation and analysis of a larger set of DBTs could provide more specific information on the regulation of many biological pathways.

Presenting Author

Pierre Martinez (pierre.martinez@kcl.ac.uk)
King's College London

Author Affiliations

1) Randall Division of Cell and Molecular Biophysics, King's College London, UK. 2) MRC Centre for Developmental Neurobiology, King's College London, UK. 3) Nestlé Research Centre, Lausanne, Switzerland.

Acknowledgements

Engineering and Physical Sciences Research Council (EPSRC - UK) Nestlé (Switzerland)

E-35. Cross-checking experimental results with publicly available gene expression data: a query-driven approach

De Smet R (1,), Hermans K (1), De Keersmaecker S (1), Marchal K (1)*

Different experimental assays highlight different aspects of the studied biology. Considering the wide availability of public microarray data an appealing approach is to interpret the results of experimental assays in light of these data. The exploration of such compendia for experimental results requires a query-driven search approach. Existing query-driven approaches, however, either fail to deal with the potential functional diversity of the experimental gene set or require the method to be run on each gene separately, resulting in a large number of solutions that needs to be evaluated.

Materials and Methods

We introduce a new computational tool which builds on an existing query-driven biclustering strategy and that given a list of genes derived from an experimental assay identifies distinct biclusters in a gene expression compendium to which the genes within the list belong. The method implements principles from consensus clustering and graph clustering to aggregate multiple query-driven biclustering results, each of them derived for a single query-gene, in post-processing. In particular different ways of assembling the query-driven biclustering results into a consensus solution are evaluated.

Results

To illustrate the effectiveness of this approach the method was applied to an Escherichia coli (ChIP-chip experiment) and Salmonella Typhimurium (single cell experiment) case study. The approach efficiently removes the redundancy amongst the query-driven biclustering results and avoids defining separate thresholds for each query-driven biclustering solution. In particular genes from the experimental assays are separated into multiple consensus biclusters, allowing for evaluation of the experimental results in terms of the genes and the conditions belonging to the bicluster.

Discussion

We illustrate that the here introduced approach might reveal different aspects of the studied biology and as such allow for a more holistic interpretation of the experimental results. In addition, it might reveal experimental inconsistencies.

Presenting Author

Riet De Smet (riet.desmet@esat.kuleuven.be)
KULeuven

Author Affiliations

(1) Department of Microbial and Molecular systems, Katholieke Universiteit Leuven, Belgium.

Acknowledgements

This work is supported by: 1) KUL: GOA AMBioRICS, GOA/08/011, CoE EF/05/007, SymBioSys; CREA/08/023 2) IWT: SBO-BioFrame; 3) IUAP P6/25 (BioMaGNet); 4) FWO IOK-B9725-G.0329.09; 6) HFSP-RGY0079/2007C.

E-36. Using tiling microarrays to predict transcription start sites: application to Lactobacillus plantarum WCFS1

Todt T (1,2,), Wels M (1,3,4), Siezen RJ (1,3,4,5), van Hijum SAFT (1,3,4,5), Kleerebezem M (3,4,5)*

Prediction of transcription start sites (TSS) in prokaryotes allows to accurately determine upstream cis regulatory promoter sequences. Microarray tiling arrays offer a cost-effective alternative to transcriptome sequencing to determine approximate TSS locations for a large complement of genes in a given organism. Our goal is to predict genome-wide TSS by combining existing gene annotation information of L. plantarum WCFS1 with whole-genome tiling microarray expression data.

Materials and Methods

In order to calibrate our algorithm, a training gene set was compiled using the following criteria: (1) gene shows expression signal in tiling array data, (2) predicted promoter location is upstream (up to 1 kb) of the gene, and (3) a predicted transcription start site is situated between predicted promoter location and gene.

Results

In this gene set of 620 genes, we found that 53% of the predicted transcription start sites are in close proximity (up to 30 nucleotides) to predicted promoter locations. For the other TSS, predicted promoter locations were further upstream or were not predicted yet.

Discussion

We are currently adapting our method to predict genome-wide TSS and to identify new promoter sequences based on these approximate TSS locations. Current results indicate that the resolution of tiling arrays limits determining exact TSS locations. Transcriptome sequencing would be a more suited means to this end.

Presenting Author

Tilman Todt (tilma.todt@han.nl)
HAN

Author Affiliations

1 Radboud University Nijmegen Medical Centre, Centre for Molecular and Biomolecular Informatics, P.O. Box 9010, 6500 GL Nijmegen, the Netherlands; 2 HAN University, P.O. Box 6960, 6503 GL Nijmegen, the Netherlands; 3 NIZO food research, P.O. Box 20, 6710 BA Ede, the Netherlands; 4 TI Food and Nutrition, P.O. Box 557, 6700 AN Wageningen, the Netherlands 5 Kluyver Centre for Genomics of Industrial Fermentation, P.O. Box 5057 2600 GA Delft, the Netherlands.

E-37. Predicting regulatory networks from large scale time-course microarray data under several stimulation conditions on MCF-7 cells

Shiraishi Y (1,), Saeki Y (1), Nagashima T (1), Yumoto N (1), Takahashi K (1), Okada M (1)*

Distinct extracellular stimulations lead to different biological regulations in many cell lines. For example, epidermal growth factor (EGF) induces proliferation whereas heregulin (HRG) induces differentiation in MCF7 human breast cancer cells. Furthermore, several researches implied that multiple signaling transduction pathways such as mitogen-activated protein kinase (MAPK) pathway and phosphoinositide 3-kinase (PI3K) -Akt pathway play important roles to determine these cell decisions.

Materials and Methods

To understand the relationships between ligands, pathways and biological regulations, we analyzed large scale time-course microarray data collected under several stimulation conditions (estrogen, EGF, HRG, MAPK inhibitors, Akt inhibitors, etc.) using MCF7 cells. First, we extracted several gene sets with similar expression patterns on many stimulation conditions using some statistical methods. Then we investigated the feature of each gene set by performing enrichment analysis of transcription factor binding motifs and gene function annotations.

Results

Our approaches found several important gene sets, to which some transcription factors known to serve significant roles on breast cancer are associated. Furthermore, we performed wet-lab analysis on some of the transcription factors identified above and confirmed that their activity levels change dynamically on corresponding stimulation conditions. Finally, using extracted gene sets and its characteristics, we constructed a putative network which describes how each ligand activates transcription factors through signaling pathways and induces biological regulations.

Discussion

Our results will be important implications for biological researchers, especially those dealing with MCF-7 cells. We believe that using gene expression patterns with some perturbations, e.g., inhibitors is a promising approach for inferring gene regulatory networks, because it will increase the variety of expression patterns and help to classify the gene sets. Still, dealing with large scale time-course microarray data under multiple conditions is not a simple task. It is necessary to develop statistical methods for taking advantage of multiple stimulation conditions further.

Presenting Author

Yuichi Shiraishi (yshira@riken.jp)
RIKEN RCAI

Author Affiliations

(1) RIKEN Research Center for Allergy and Immunology

Acknowledgements

Grant-in-Aid for Young Scientists (B) 21700316 of the Ministry of Education, Culture, Sports, Science and Technology.

E-38. Inferring regulatory networks from expression data using tree-based methods

Huynh-Thu VA (1,2,), Irrthum A (1,2), Wehenkel L (1,2), Geurts P (1,2)*

One of the pressing open problems of computational systems biology is the elucidation of the topology of genetic regulatory networks (GRNs) using high throughput genomic data, in particular microarray gene expression data. The Dialogue for Reverse Engineering Assessments and Methods (DREAM) challenge aims to evaluate the success of GRN inference algorithms on benchmarks of simulated data. In this article, we present a new algorithm for the inference of GRNs that was best performer in the DREAM4 In Silico Multifactorial challenge.

Materials and Methods

Our GRN inference algorithm decomposes the prediction of a regulatory network between p genes into p different regression problems. In each of the regression problems, the expression pattern of one of the genes (target gene) is predicted from the expression patterns of all the other genes (input genes), using tree-based ensemble methods. The importance of an input gene in the prediction of the target gene expression is taken as an indication of a putative regulatory link. The whole network is then reconstructed by aggregating putative links over all genes.

Results

We applied our algorithm on the DREAM4 In Silico Multifactorial challenge simulated data. The goal of this challenge was to infer five networks of 100 genes from static steady-state expression profiles resulting from slight perturbations of all genes. Our method got the best performance among twelve challengers. Experiments were also carried out using 907 Escherichia coli Affymetrix arrays and 3433 known E. Coli regulatory interactions from RegulonDB. They show that our algorithm compares favorably with existing algorithms to decipher the genetic regulatory network of Escherichia coli.

Discussion

In conclusion, we propose a new algorithm for GRN inference that performs well on both synthetic and real gene expression data. The algorithm, based on feature selection with tree-based ensemble methods, does not make any assumption about the nature of gene regulation, can deal with combinatorial and non-linear interactions, produces directed GRNs, and is fast and scalable. It is furthermore simple and generic, making it adaptable to other types of genomic data and interactions.

URL

http://www.montefiore.ulg.ac.be/~huynh-thu/software.html

Presenting Author

Vân Anh Huynh-Thu (vahuynh@ulg.ac.be)
University of Liège

Author Affiliations

(1) Department of Electrical Engineering and Computer Science, Systems and Modeling, University of Liège, Liège, Belgium (2) GIGA-Research, Bioinformatics and Modeling, University of Liège, Liège, Belgium

Acknowledgements

This work was partially funded by the Interuniversity Attraction Poles Programme (IAP P6/25 BIOMAGNET), initiated by the Belgian State, Science Policy Office and by the European Network of Excellence PASCAL2. VAHT is recipient of a F.R.I.A. fellowship and PG is Research Associate of the F.R.S.-FNRS, Belgium.

E-39. Dynamic profile of transcription factors inferred from mRNA expression time courses in Gefitinib-treated lung cancer cells

Nagao H (1,), Yoshida R (1), Saito M-M (1), Imoto S (2), Nagasaki M (2), Yamaguchi R (2), Yamauchi M (2), Goto N (2), Miyano S (2), Higuchi T (1)*

Transcription regulatory pathways respond to stimuli by changing activities of transcription factors (TFs) and enhancing rates of gene decoding. DNA microarray experiments enable us to measure transcript levels of mRNA molecules that reflect underlying activity of TFs. Our study builds on a reverse-engineering system to reconstruct temporal profiles of activating TFs using given mRNA time courses. The proposed method is tested on lung cancer cells, clarifying differences in TFs involving pathways between Gefitinib-sensitive/resistive lung tumors treated with epidermal growth factor (EGF).

Materials and Methods

We measure gene expression time courses of Gefitinib-sensitive/resistive lung tumor cells (PC9 and PC9GRM2), treated with EGF and/or Gefitinib. To understand differences in TFs involving pathways between the different types of cells, for each of the experimental conditions, the proposed method reconstructs temporal profiles of activating TFs called "TF dynamic maps" by statistically evaluating significances of synchronization in temporal gene expression patters involving a particular TF. The list of TFs and their target genes were retrieved from TRANSPATH database.

Results

By comparisons of the obtained TF dynamic maps, we received preliminary results as follows; (1) key TFs on Gefitinib-response pathways were clearly identified from PC9 profiles, (2) the TF dynamic maps uncovered sets of synchronizing TFs dynamically controlled during propagations of EGF and/or Gefitinib stimuli, (3) PC9GRM2 showed significantly lower responses than PC9 in both EGF and EGF+Gefitinib treated cells, (4) Activity levels of most transcriptional activators and repressors decreased and increased respectively at the almost same timing (around 20 hours after stimulation) in PC9GRM2.

Discussion

The constructed TF dynamic maps are utilized in understanding which TFs play significant regulatory roles in EGF and/or Gefitinib treated tumor cells and how the Gefitinib acts on TFs dynamically. More interestingly, conduction of a more detailed analysis can distinguish controllable and uncontrollable TFs that are able/unable to induce significant changes in gene regulatory systems by conducting intervention experiments. A list of controllable and uncontrollable TFs can be vital information in design of molecular-targeted drug.

Presenting Author

Hiromichi Nagao (hnagao@ism.ac.jp)
The Institute of Statistical Mathematics

Author Affiliations

1. The Institute of Statistical Mathematics 2. The Institute of Medical Science, the University of Tokyo

Acknowledgements

This research is supported by the next-generation supercomputer project of RIKEN.

E-40. Deciphering the genomic plasticity of bacterial species

Janky R (1,), Harvill ET (2), Babu MM (1)*

The genomes of bacterial species show enormous plasticity in the function of individual genes, reflected by a rapid evolution in genome organization and regulatory interactions. One of the major challenges is to decipher the mechanisms of evolution of gene regulation which leads to bacterial adaptation. We developed a computational procedure and compared the predicted regulatory systems from four closely related bacterial species in the genus Bordetellae, of which three are pathogenic with different host specificity.

Materials and Methods

We used a phylogenetic footprint discovery method to detect conserved regulatory motifs in promoter regions of Beta-proteobacteria. To study the regulatory motifs evolution within Bordetella species, we defined a conservation profile according to its presence or not in their promoters. The distribution of predicted motifs in different phylogenetic profile was compared to those from random selection of Bordetella promoters.

Results

We found that the distribution of predicted motifs on the Bordetella promoter sequences is not random. Cis-regulatory motifs matching pathogenic sequences are significantly over-represented, while others detected in any combination of non-pathogenic and pathogenic promoter sequences are significantly under-represented. Two case studies with enriched pathogenic regulations will be discussed.

Discussion

Such comparative approach at the level of regulatory analysis may allow us to predict pathogen-specific regulatory interactions and provide new insights into bacterial adaptation due to the evolution of the regulation. We will apply this approach to other taxonomic groups with pathogenic species.

Presenting Author

Rekin's Janky (rjanky@mrc-lmb.cam.ac.uk)
Medical Research Council - Laboratory of Molecular Biology.

Author Affiliations

(1) Structural Studies Division. Medical Research Council - Laboratory of Molecular Biology. Hills Road, Cambridge, CB2 0QH. United Kingdom. (2) Department of Veterinary and Biomedical Science. The Pennsylvania State University, USA.

Acknowledgements

M.M.B. and R.J. acknowledge the Medical Research Council for financial support. R.J. was also supported by a post-doctoral grant from the Wiener-Anspach Foundation (Belgium). M.M.B. thanks Darwin College, Cambridge and Schlumberger Ltd. for generous support.

E-41. Expression arrays & outlying processes: building a database of outliers

Berger F (1,), Kroll KM (1), van Dorp M (1), Carlon E (1)*

Expression arrays are used to study the expression profile of the whole genome. Focusing on Affymetrix expression arrays, Kroll M. & Carlon E. recently introduced affyILM, a package estimating transcript concentrations. Their method makes use of hybridization free energies, extended Langmuir isotherm, and physical features of neighboring spots. In contrast to other methods, the output is target concentrations measured in pM (picoMolar). Based on this approach, we aim to identify probes that are systematically detected as outliers, as well as to investigate the associated outlying processes.

Materials and Methods

To develop the database, we used the MySQL database management system. The interface has been developped using the php scripting language, and is hosted on the apache http server. Datasets were retrieved from the EBI ArrayExpress repository (accession ids : E-GEOD-994 and E-TABM-145). The analyses were performed using R, Bioconductor and the affyILM package. Sequences alignments were performed using the NCBI blastn methodology (short sequences algorithm), against GenBank and RefSeq human nucleotide sequences.

Results

We build an exploratory database of probes features, focusing on the HG-U133A chip model. Theoretical outliers are retrieved from previous publications, and from alignment results between probes and targets. Empirical detection of outliers is performed assuming that concentration values within a probe-set follow the normal distribution. A php-interface compares the concentrations associated to transcript-specific probes from publicly available datasets. By means of our web interface, the database can be used in combination with public datasets to track the behavior of outlying probes.

Discussion

The outlier database and the web interface will be used to highlight the outlying processes that are involved. In the future, we plan to modify affyILM to improve the concentration estimation of probes associated to specific outlying processes (appropriate estimation of hybridization free energies). In addition, the current database will be enlarged with other Affymetrix expression arrays.

Presenting Author

Fabrice Berger (fabrice.berger@fys.kuleuven.be)
KULeuven

Author Affiliations

(1)Institute for Theoretical Physics, KULeuven

E-42. Detecting significantly correlated gene clusters and their interactions in a continuous scale-space

Van Dyk E (1,2,), Reinders M (2), Wessels L (1)*

Biological complexity stems from the large number of possible interactions between molecular entities occurring at different levels of modularity. Low-level (i.e. gene-gene) interactions include co-operating interactions and mutually exclusive interactions. Examples of higher-level interactions are module (collection of genes) interactions. Like single genes, modules can be activated or repressed. Modules can also be mutually exclusively and co-occurrently activated. These gene or module level interactions could reveal critical steps in tumorigenesis.

Materials and Methods

We are mainly interested in defining gene modules on a continuous scale space using kernel convolution methods applied to microarray gene expression data from breast-cancer tumors. We map human genes on their genomic positions and create a correlation matrix where each entry represents the Pearson correlation of a gene pair over all samples. In this newly defined space, we perform kernel smoothing (with different kernel widths) to reveal significant clusters at different scales. We compensate for the non-uniformity of the gene positioning in this space to maximize the power of significance.

Results

We present theoretical methods for finding statistically significant clusters in the kernel-convoluted space (non-randomly occurring clusters). These include methods for minimizing false-negative rates (chance of missing clusters) and approaches for controlling the FWE rate (false-positives). We compute a significance threshold that in general depends on the position in space and thus compensates for the non-uniform spacing of genes. We demonstrate this approach on two breast cancer gene expression data sets obtained from different micro-array platforms (Agilent and Affymetrix respectively).

Discussion

We focus on finding clusters of genes that are significantly correlated. These clusters can be defined as modules of genes that are close together on the genome and are identified by significant peaks on the diagonal of the smoothed correlation matrix. Clusters far from the diagonal represent interactions between modules (such as co-occurring or mutually exclusive). This also allows for the discovery of modules at different scales and inter-scale interactions. Identification of stable modules in different data sets can be integral for providing robust feature extraction tools.

Presenting Author

Ewald van Dyk (e.v.dijk@nki.nl)
The Netherlands Cancer Institute

Author Affiliations

1 Bioinformatics and statistics, Division of Molecular Biology, The Netherlands Cancer Institute, Amsterdam, The Netherlands 2 Delft Bioinformatics Laboratory, Faculty of EEMCS, Delft University of Technology, Delft, The Netherlands

E-43. MAMMOD: Multi-Agent Motif and MOdule Discovery

Van Delm W (1,2,), Ayoubi T (2), De Moor B (1), Moreau Y (1)*

Deciphering non-coding DNA is a challenging task. The first generation of tools searched for binding sites of a single transcription factor (TF) at a time. A thorough community assessment in 2004 however revealed that much work remained to be done. Gradually techniques became more advanced, looking for clusters of binding sites of a collection of TFs. Very recently, the flood of new methods was accompanied by new test benches to validate and compare alternatives. Despite the improvements, many predictions still appear to be false positives.

Materials and Methods

We adopt a Markov random field (MRF) that connects the unknown variables with data. Data of sequence conservation and protein interaction is included. The configuration of bound TFs is modeled with a semi-hidden Markov model. The other components of the MRF are classical statistical distributions, of which the parameters are fitted with expectation-maximization methods. Samples are drawn from the posterior distribution with a novel sampling scheme, that generalizes a grouped Gibbs sampling algorithm. We use Friedman tests to statistically compare performance measures of different algorithms.

Results

By using both simulated and benchmark data sets of skeletal muscle and liver tissue, we demonstrate that our novel algorithm for Bayesian inference can compete with several state-of-the-art methods in predicting relevant transcription factors and cis-regulatory modules.

Discussion

We have introduced a general computational framework for Bayesian inference of cis-regulatory motifs and modules in eukaryotic promoters. The framework allows the integration of alternative algorithms and data sources. Further investigation of the algorithmic behavior is necessary to control better prediction specificity, robustness to noise and improper modeling, and especially computation time.

Presenting Author

Wouter M.H.J. Van Delm (wvandelm@gmail.com)
VIB - MicroArray Facility

Author Affiliations

(1) Dept. Electrical Engineering ESAT-SDC, Katholieke Universiteit Leuven, (2) MicroArray Facility, VIB

Acknowledgements

This research is funded by a Ph.D. grant of the Institute for the Promotion of Innovation through Science and Technology in Flanders (IWT-Vlaanderen). We are grateful to the Department of Electrical Engineering (ESAT-SCD) of the Katholieke Universiteit Leuven and the Flemish Institute for Biotechnology (VIB) for their scientific support.

E-44. Modulon identification from bacterial gene expression array data using an improved supervised approach

Hedge S (1), Permina E (2,), Medvedeva Y (2), Mande S C (1), Makeev V (1,3)*

Inferring regulatory pathways from microarray data is an important preparatory stage for system biology modeling of a bacterial cell. A variety of methods have been developed to solve this problem. Using a traditional approach, researchers focus on finding clusters of correlated expression values over the set of various conditions. This leaves us with the major problem of improving the quality of resulting data

Materials and Methods

We suggest a strategy dealing with microarray data, starting with identifying noise component in expression array, exploiting the idea that pairs of genes belonging to the same operon make a natural set of positively correlated pairs of genes. Negatively correlated genes observed in the set of operon pairs allow calculating the noise component. The next step is identifying clusters of coexpressed genes starting from a set of known genes which are divided into two subsets representing the essential and optional parts of the modulon in order to make the approach more flexible

Results

We have develop an easy-to-handle software for calculating the cliques of potentially coexpressed genes and have applied our technique to LexA, PurR, and heat shock (Sig32) operons in E. coli, verifying questionable regulon members and predicting the new ones. We succeeded in prediction of additional modulon participants and have compared the results obtained from several independent datasets. Our results are consistent with the available experimental data

Discussion

The work demonstrates the importance of using reference sets for which the correlation data can be calculated a priori and the advantage of supervised method for dealing with microarray data. The use of structured set of known genes allows to get clusters of potentially coexpressed genes which are more compact and reasonable from the biological point of view. Coordinated expressions of genes belonging to different regulons renders possible revealing the global structure of bacterial regulatory network.

Presenting Author

Elizabeth A. Permina (enelkinsan@gmail.com)
GosNIIGenetika

Author Affiliations

(1) Centre for DNA Fingerprinting and Diagnostics, India; (2) State Research Institute for Genetics and Selection of Industrial Microorganims, GosNIIgenetika, Russia; (3) Engelhardt Institute of Molecular Biology, Russia

Acknowledgements

This work was partially supported by Russian Fund of Basic Research Project 08-04-91318 and Russian President Young Scientist Grant MK-4805.2009.4

E-45. GCQM: quality assessment of microarray studies and samples using gene correlations

Venet D (1,), Detours V (2), Bersini H (1)*

A large amount of high-throughput gene expression data is available in public repositories. However, the quality of those data can vary dramatically from platform to platform, study to study and sample to sample. Determining data quality is of the utmost importance. Current quality measures are platform-specific, and cannot be used to compare studies. We propose here a biology-motivated quality measure, the Gene Correlation Quality Measure (GCQM), based on the consistency of gene-gene correlations.

Materials and Methods

GCQM allows for the estimation of qualities of both studies and samples. First, the average of all gene-gene correlation matrices in our database was calculated. Then, the quality of studies is obtained by comparing its gene-gene correlation matrix with the expected correlations. The relative quality of a sample is obtained by comparing the quality of its study with and without that sample.

Results

the quality of studies, samples and platforms. This allowed to discover serious issues with three studies. We show that qualities of both studies and samples are linked to the statistical significance of the biological results, hence higher quality studies have a higher signal/noise ratio. We show that GCQM does give a higher quality to biological outliers than warranted, so that those will not be discarded. We also determined an average quality for many platforms.

Discussion

It could come as a surprise that average gene-gene correlations allow for the determination of data quality. It works because gene-gene correlations have a distribution which is meaningful for any study, being largely based on genes that belong to the main cellular processes (e.g. respiration, cell cycle, proliferation). Biological variability slightly up- or down-regulates those processes in every sample, creating those correlations. The quality of biological outliers is over-valued because GCQM is a signal/noise measure.

Presenting Author

David Venet (davenet@ulb.ac.be)
Université Libre de Bruxelles

Author Affiliations

1 IRIDIA, Université Libre de Bruxelles, 50 Av. F. Roosevelt, CP 194/6, B-1050 Brussels, Belgium 2 IRIBHM, Université Libre de Bruxelles, Campus Hospitalo-Facultaire Erasme, 808 route de Lennik, B-1070 Brussels, Belgium

Acknowledgements

This work was funded by the ICT Impulse program 2006, Brussels Capital Region, Belgium

E-46. Integration of correlation structure improves performance of pathway-based classification

*Cho SB (1, *), Kim JH (2,3)*

In a pathway-based classification analysis, considering the correlation structure between genes is important. We have developed a boosting kernel density Bayesian classifier (boostKDBC) algorithm to capture the correlation structure during classification.

Materials and Methods

The pathway expression profiles from four breast cancer data sets and simulation data were tested to validate the boostKDBC method. Multivariate kernel density estimation was plugged in boosting procedure.

Results

In real data analysis, the boostKDBC algorithm outper-formed other benchmark algorithms. Simulation study indicated that the performance of the boostKDBC algorithm was superior to other algorithms with data having a gene–gene correlation structure. The pathways listed in the results provide many invaluable hypotheses on the pathophysiologic mechanism of breast cancer.

Discussion

It seeemed that boosting reinfored accuracy of conditional density estimation, which led to better perforamnce than any other benchmark algorithms. Based on our results, we assume that the boostKDBC algorithm is appropriate for pathway-based classifica-tion analysis.

Presenting Author

SungBum Cho (sbcho@korea.kr)
center for disease control, korea

Author Affiliations

(1) Division of biomedical informatics, National Institute of Health, Korea, Seoul 122-401. (2) Seoul National University Biomedical Informatics (SNUBI) (3) Division of Biomedical Informatics, Seoul Na-tional University College of Medicine, Seoul 110-799, Korea

E-47. Multi Experiment Matrix: web tool for mining co-expressed genes over hundreds of datasets

Adler P (1,), Kolde R (2,3), Kull M (2,3), Peterson H (1,3), Reimand J (2), Tkatšenko A (2,3), Vilo J (2,3)*

Accumulation of gene expression data in public domain has raised the opportunity to discover new facts by re-analyzing the existing experiments. Co-expression over many gene expression datasets has been proven useful in many areas of molecular biology and bioinformatics, such as network reconstruction and gene function prediction. However, the tools for this are lacking. Here, we present MEM, a web-based resource for finding co-expressed genes over large collections of public microarray experiments.

Materials and Methods

The search goes over hundreds of gene expression datasets from ArrayExpress covering wide spectrum of biological conditions from stem cells to different diseases and cancer. The essence of MEM is a novel rank aggregation method that combines similarity searches across individual data sets into a global ordering by selecting automatically the appropriate experiments and assigning a combined p-value for the significance of the similarity across all data. For more specific queries we have various mechanisms for dataset selection.

Results

As a result we have developed a web-based tool to query over many hundreds of gene expression datasets to find similarly expressed genes. Using known transcription factor NANOG as an example gene, First we demonstrate how we are able to identify relevant datasets for query. Then looking for known targets of NANOG among top similarly expressed genes in each individual dataset separately and using MEM, we show that using our rank-aggregation schema to combine rankings from each dataset improves the overall yield.

Discussion

We have developed a tool for finding co-expressed genes over many public gene expression datasets. The tool has many possible applications, such as gene function prediction or network reconstruction. The search is powered by a novel rank aggregation algorithm. User interface is highly interactive and provides additional information about the query and the results.

URL

http://biit.cs.ut.ee/mem

Presenting Author

Priit Adler (adler@ut.ee)
University of Tartu, IMCB

Author Affiliations

(1) Institute of Molecular and Cell Biology, University of Tartu (2) Institute of Computer Science, University of Tartu (3) Quretec Inc

Acknowledgements

This work has been supported by EU FP6 ENFIN, FunGenES, Estonian Science Foundation 5724, 7437

E-48. Linking parallel measurements of high-throughput miRNA, gene and protein expression data

Gade S (1,), Binder H (2), Beißbarth T (3)*

The parallel measurement of different types of high throughput expression data, like gene, miRNA and protein data, gives rise to new tasks of data integration, specifically, when also the combined effect on a clinical endpoint is of interest. Different strategies have been formulated, performing the integration on different levels of the analysis process. Specifically, miRNAs are small non-coding RNAs which play an important role during cancer development. The details of the miRNA mediated regulation mechanism are not fully understood. Though it is known that the binding of a miRNA to its target transcript can lead to degradation and consequently to a measurable change in the level of the mRNA.

Materials and Methods

We proposed a graph-based work flow to integrate miRNA and gene expression data. By combining the correlation structure of the two data sets with target predictions we were able to construct a bipartite graph with connections between miRNAs and genes. Meta information were added and used to distinguish reasonable sub clusters in this graph. This structure was used to perform an over-representation analysis of targets and effected pathways. Furthermore, the bipartite graph is interpretable as a pathway linking different kinds of features. Similar to gene pathways it is suitable as an additional knowledge for classification tasks.

Results

We applied this method to a colorectal cancer data set consisting of 97 samples and measurements of miRNA and gene expressions. With help of the graph structure we were able to identify clusters of miRNAs with significant correlations to genes in cancer related pathways like cytokine-cytokine receptor interaction and the Jak-STAT signaling. Finally, for linking the miRNA and gene expression data to a clinical endpoint, a componentwise boosting approach was employed that incorporates the graph information for fitting a multivariable model and selects a small set of prognostic molecular entities.

Discussion

Combining different kind of omics data is a challenging task. With this graph based workflow we presented a method that combines the different data sets in a graph model giving rise to e.g. pathway over representation analysis and classification methods for predicting clinical endpoints. Thereby, not only the features of both data sets but the coherences between these features are used.

Presenting Author

Stephan Gade (s.gade@dkfz-heidelberg.de)
German Cancer Research Center Heidelberg

Author Affiliations

(1) German Cancer Research Center Heidelberg (2) University of Freiburg (3) University of Göttingen

E-49. An extended translational network involving transcription factors and RNA binding proteins derived fro, the identification of phylogenetic footprints in 5' and 3' untranslated regions

Dassi E(1,), Tebaldi T(1), Zuccotti P(2), Riva P(2), Quattrone A(1)*

Sequence motifs found in the untranslated regions (UTRs) of mRNAs could be cis functional element involved in the post-transcriptional regulation of gene expression, given that most RNA-binding proteins (RBPs) and non coding RNAs (ncRNAs) bind to target sequences in either the mRNA 5' or 3' UTRs to influence the fate of a transcript. If acting in evolutionarily conserved control networks, these motifs would be subjected to negative (purifying) selection, so that they could be identified by extracting orthologous sequences in 5’ and 3’ UTRs for their high similarity among a variety of species.

Materials and Methods

To map hyperconserved elements in the 5’ and 3’ UTR we exploited the UCSC genome-wide alignment of 46 vertebrate species. Hyperconservation measure was derived both from sequence conservation and extent of the phylogenetic tree covered by the UTR alignment, each parameter weighting equally. Hyperconserved elements (hyp-SUBs) were identified by searching in all UTRs for fully conserved seeds that were then extended upstream and downstream until the score did not fall under the 85% threshold. Hyp-SUBs were then characterized by ontologies, motifs and regulatory factors binding sites search.

Results

Analysis of hyp-SUBs resulted in the observation of no matching between them and long ncRNAs or microRNA seed targets. Hyp-5’SUBs were enriched for HOX gene family members (suggesting the presence of an unappreciated 5’UTR-targeted mechanism of HOX gene regulation) and for kinase family genes; hyp-3’SUBs are enriched in genes coding for RBPs of the RRM type, in line with experimental findings indicating a tendency for RBP expression to be regulated post-transcriptionally. A benchmark experiment will be presented to identify the mechanism exploiting these hyp-3'SUBs for the regulation of RBPs.

Discussion

As a proof of validity of our approach, we identified a subset of hyp-3'-SUBs belonging to the 3'UTR of histones mRNAs: a motif search in these hyp-SUBs revealed that most of them contains the binding motif of the histone stem loop binding protein (SLBP), which binds to the 3' histones UTR to stabilize their mRNAs. This protein and its binding sites are very conserved across evolution, indicating that our method can identify truly conserved elements in UTRs. Also, a network of RBPs controlled by an upstream master RBPs has been identified and partially experimentally confirmed in the same way.

Presenting Author

Erik Dassi (dassi@science.unitn.it)
CIBIO - University of Trento

Author Affiliations

(1) Laboratory of Translational Genomics, Centre for Integrative Biology, University of Trento (2) Department of Biology and Genetics for Medical Sciences, University of Milan

E-50. Unvealing heterogeneity of stage II and stage III colorectal cancer by defining molecular subtypes

Budinska E (1,2,3,), Popovici V (1), Delorenzi M (1,4)*

Despite the continuous efforts for characterization of the colorectal cancer, a large heterogeneity in survival experience is still observed. Furthermore,the treatment efficacity varies widely even inside the subpopulations defined by clinical and mutational status variables. Our interest was to provide a complementary characterization of stage II and III patients, built in a bottom-up fashion, that would allow further stratification of the patient population, better outcome prediction and development of more effective targeted treatments.

Materials and Methods

Stage II and III CRC samples from three public datasets were considered for analysis. The unsupervised consensus clustering of the patients led to the the definition of an optimal number of subgroups. Gene modules defining the subgroups of patients were defined by hierarchical clustering. The robustness of the patient subgroups was assessed across datasets by cluster reproducibility measures on the gene modules. The subgroups were characterized by the pathway analysis, the comparison with multiple normal and cancer cell types and the assessment of clinical parameters.

Results

Using techniques of meta-analysis we have defined four major molecular subtypes of stage II and stage III colorectal cancer, reproducible across the three public datasets. The pathway analysis indicated differences in major cellular proceses among subgroups of patients, suggesting a novel perspective on the classification of the colorectal cancer. The identified subgroups remain, however, heterogeneous in terms of clinical parameters, idicating a complementarity between the gene expression profiles and clinical characteristics.

Discussion

We have identified four major molecular subtypes of colorectal cancer, which are highly robust across studies. The most important outcome consists in the observation that from a clinical perspective the subgroups remain heterogeneous. This suggests that the newly proposed stratification should be considered as a complementary characterization of colorectal cancer. The limited information does not allow deeper analysis of treatment implications on the subgroups. Nevertheless, we are confident that the proposed classification represents true biological differences between patients.

Presenting Author

Eva Budinska (eva.budinska@isb-sib.ch)
Swiss Institute of Bioinformatics

Author Affiliations

1 Swiss Institute of Bioinformatics, Lausanne, Switzerland 2 Institute of Biostatistics and Analyses, Brno, Czech Republic 3 Masaryk Memorial Center Institute, Brno, Czech Republic 4 Department of Research, Lausanne University Hospital, Lausanne, Switzerland

E-51. A method for improved prediction of in vivo transcription factor binding sites by using biophysical DNA features

Hooghe B (1,2,3, +), Broos S (1,2,3,, +) ,Van Roy F (2,3), De Bleser P (1,2,3)*

Transcription factor binding sites (TFBSs) are short (6-20 bp) DNA sequences that can be regarded as base units of transcription regulation. The binding interactions with their protein partners, transcription factors (TFs), determine a big part of the spatiotemporal gene expression patterns. In silico identification of these basic units has in principal a huge potential to speed up research of gene regulation. The accurate in silico identification of TFBSs that do really bind in vivo and not just in vitro has however proven to be a difficult task.

Materials and Methods

Prediction by PWMs was used for the initial alignment of true BSs and in the first step of TFBS identification. PWMs and the search algorithm MATCH were from TRANSFAC version 2008.4. We linked all biophysical values to tetra- or pentanucleotides to capture the sequence-dependent biophysical variation caused by neighboring nucleotides. Most values were collected from literature, others are own calculations. We used 3DNA to calculate this for several models of both unbound DNA and either protein-bound or nucleosome-bound DNA.

Results

We then compared our approach to currently employed sequence-based methods that use only nucleotide sequence information. To ensure that methods and not training data sets were being compared, we trained these models with sequences sampled from the same data set we used to train our model. For each TF, we compared the best biophysical model to own constructed PWMs and the TRANSFAC PWMs for that TF: our approach performed significantly better for all 20 TFs.

Discussion

The reported improvements in accuracy make a huge difference upon identification when identifying TFBSs. We took one trained biophysical model of Stat1 and used it to search binding sites (BSs) for this TF in the sequences of the corresponding positive test set. We also used the TRANSFAC PWM for Stat1 (V$STAT1_01) to identify BSs in these sequences. This PWM is quite qualitative and the difference of accuracy between the PWM and the best biophysical model is the smallest of all TFs shown. Yet the biophysical model is clearly the best choice for reliable identification of in vivo TFBSs.

Presenting Author

Stefan Broos (stefan.broos@dmbr.vib-ugent.be)
DMBR, VIB - Ghent University

Author Affiliations

1. Bioinformatics Core Facility, VIB, B-9052 Ghent, Belgium 2. Department of Biomedical Molecular Biology, Ghent University, B-9052 Ghent, Belgium 3. Department for Molecular Biomedical Research, VIB, B-9052 Ghent, Belgium

Acknowledgements

This work was made possible by Flanders Institute for Biotechnology (VIB) and funding from the Agency for Innovation through Science and Technology in Flanders (IWT).

E-52. Nonlinear dimension reduction and clustering by minimum curvilinearity unfold neuropathic pain and tissue embryological classes

Cannistraci CV (1,2,3,4,5,), Ravasi T (1,5), Montevecchi FM (3), Ideker T (5), Alessio M (2)*

Nonlinear small datasets, which are characterized by low numbers of samples and very high numbers of measures, occur frequently in computational biology, and pose problems in their investigation. Unsupervised Hybrid-two-phase procedures (H2P) - specifically dimension reduction (DR) coupled with clustering - provide valuable assistance, not only for unsupervised data classification, but also for visualization of the patterns hidden in highdimensional feature space.

Materials and Methods

'Minimum Curvilinearity' (MC) is a principle that - for small datasets - suggests the approximation of curvilinear sample distances in the feature space by pairwise distances over their minimum spanning tree (MST), and thus avoids the introduction of any tuning parameter. MC is used to design two novel forms of nonlinear machine learning (NML): Minimum Curvilinear Embedding (MCE) for DR, and Minimum Curvilinear Affinity Propagation (MCAP) for clustering.

Results

Compared with several other unsupervised and supervised algorithms, MCE and MCAP, whether individually or combined in H2P, overcome the limits of classical approaches. High performance was attained in the visualization and classification of: 1) pain patients (proteomic measurements) in peripheral neuropathy; 2) human organ tissues (genomic transcription factor measurements) on the basis of their embryological origin.

Discussion

MC provides a valuable framework to estimate nonlinear distances in small datasets. Its extension to large datasets is prefigured for novel NMLs. Classification of neuropathic pain by proteomic profiles offers new insights for future molecular and systems biology characterization of pain. Improvements in tissue embryological classification refine results obtained in an earlier study, and suggest a possible reinterpretation of skin attribution as mesodermal.

URL

https://sites.google.com/site/carlovittoriocannistraci/

Presenting Author

Carlo Vittorio Cannistraci (kalokagathos.agon@gmail.com)
Red Sea Integrative Systems Biology Lab, Computational Bioscience Research Center, Division of Chemical & Life Sciences and Engineering, King Abdullah University for Science and Technology (KAUST)

Author Affiliations

(1) Red Sea Integrative Systems Biology Lab, Computational Bioscience Research Center, Division of Chemical & Life Sciences and Engineering, King Abdullah University for Science and Technology (KAUST), Jeddah, Kingdom of Saudi Arabia (2) Proteome Biochemistry, San Raffaele Scientific Institute, via Olgettina 58, 20132 Milan, Italy (3) Department of Mechanics, Politecnico di Torino, c/so Duca degli Abruzzi 24, 10129 Turin, Italy (4) CMP Group, Microsoft Research, Politecnico di Torino, c/so Duca degli Abruzzi 24, 10129 Turin, Italy (5) Department of Bioengineering and Department of Medicine, University of California, San Diego, 9500 Gilman Drive, La Jolla, CA 92093 USA

Acknowledgements

We thank Ewa Aurelia Miendlarzewska for her generous assistance and for language revision and Sven Bergmann for precious suggestions. Funding: This work was supported by Fondazione CARIPLO (NOBEL GuARD Project) and by MoH RF-FSR-2007-637144. CVC was supported by the Italian Interpolytechnic School of Doctorate SIPD (http://sipd.polito.it/). TR and CVC was supported by the MH062261grant from the US National Institute of Mental Health and the King Abdullah University of Science and Technology.

E-53. Seeded cis-regulatory module discovery using rank based motif scoring

Macintyre G (1,2), Bailey J (1,2), Kowalczyk A (2,1), Haviv I (3,4,5)*

Transcription factors (TFs) act in concert by binding to regions of DNA known as cis-regulatory modules (CRMs), usually in proximity to the genes they regulate. These cooperating transcription factors play a major role in the molecular mechanisms responsible for complex gene regulatory programs observed in humans. Limited experimental mapping of CRMs in the human genome means that in silico CRM discovery methods continue to play an important role in understanding complex gene regulatory programs. We present a novel and improved in silico cis-regulatory module detection algorithm called SCRM,

Materials and Methods

Our approach does not use species conservation information to limit the CRM search space. This allows for the discovery of species specific cis-regulatory elements. In addition, we have developed a `seeded' search strategy for de novo discovery of interacting TFs, eliminating the need for prior knowledge of interacting TFs and their position weight matrices. The `seeded' search strategy is also used to determine CRM locations using a rank based scoring system that favors clusters of interacting TF binding sites around the `seed' TF binding site.

Results

Our method is designed to uncover the regulatory mechanisms involved in a particular molecular pathway for which there is a set of putatively co-regulated genes and prior knowledge of a single TF (the `seed') that may regulate those genes. For example, we provide the location and composition of CRMs involved in the response of Hela cells to IFN-gamma for which STAT1 plays a major role.We validate our predictions in vivo using a mouse c-jun knockout model demonstrating a significant loss in expression of those genes predicted to be regulated by c-jun.

Discussion

The ability of transcription factors to bind specific locations in the genome and control the transcription of a gene is typically governed by four main factors: DNA site recognition, the gene's regulatory domain, chromatin structure and co-factor interactions. We use knowledge of each of these in the development of an in silico cis-regulatory module (CRM) detection algorithm (called SCRM) for aid in interpretation of gene regulatory networks in humans. Our biologically motived design and inclusion of a single known regulating TF results in improved performance over existing CRM predictors.

Presenting Author

Geoff J Macintyre (gmaci@csse.unimelb.edu.au)
Department of Computer Science and Software Engineering, University of Melbourne

Author Affiliations

(1)Department of Computer Science and Software Engineering, The University of Melbourne, Victoria, Australia (2) NICTA, Victoria Research Lab, The University of Melbourne, Victoria, Australia (3) Bioinformatics and Systems Integration, The Blood and DNA Profiling Facility, Baker IDI Heart and Diabetes Institute, 75 Commercial Rd, Prahran, Victoria, Australia (4) Metastasis Research Lab, Peter MacCallum Cancer Centre, St Andrews Place, East Melbourne, Victoria, Australia (5) Department of Biochemistry and Molecular Biology. The University of Melbourne, Victoria, Australia

Acknowledgements

IH is supported by the following grants: [586649 NHMRC], [BC084667 US DoD BCRP], [566882 Cancer Australia], [BCTR0707358 Komen for the Cure], and [509292 NBCF]. This work was partially supported by NICTA. NICTA is funded by the Australian Government's Department of Communications, Information Technology and the Arts, the Australian Research Council through Backing Australia's Ability, and the ICT Centre of Excellence programs. GM is supported by an Australian Postgraduate Award and a NICTA Research Project Award.

E-54. Proximity-based cis-regulatory module detection using constraint programming for itemset mining

Guns T (1,), Sun H (2), Nijssen S (1), Sanchez-Rodriguez A (2), De Raedt L (1), Marchal K (2)*

cis-regulatory modules (CRMs) are combinations of Transcription Factor Binding Sites involved in the regulation of genes. The detection of CRMs is key in developing a better understanding of gene regulation. Identifying significant combinations of binding sites is a difficult computational problem. Existing techniques use heuristic methods or strong restrictions to make the problem tractable, and are not very extendible. We present an extendible technique for enumerating all potential CRMs using only a biologically well-motivated restriction on the proximity of the binding sites involved.

Materials and Methods

Our method consists of 3 phases: First, the genomic sequences under investigation are screened using an existing library of motif models, namely, position weight matrices. This screening identifies Transcription Factor Binding Site (TFBS) hits. Secondly, we use constraint programming for itemset mining to efficiently enumerate all combinations of TFBS hits that co-occur within a pre-defined distance, while avoiding undesirable redundancies. This results in an exhaustive list of potential CRMs. Lastly, the CRMs are ranked using statistical methods.

Results

We use two previously published datasets which are known to contain a true CRM composed of 3 transcription factors (TFs). One is a synthetic dataset consisting of 22 genomic sequences, the other a ChIP-seq dataset of 25 sequences. We test the scalability of our method by screening an increasing number of TFs, sampled out of 516 TRANSFAC vertebrate TFs. Our method finds all CRMs in up to 40 TFs and possibly even more. We are still in the process of comparing the sensitivity and precision/recall of our method, including the statistical ranking phase, with other tools.

Discussion

To make enumerating all potential CRMs feasible, we take special care in avoiding solutions that are equivalent to others. We believe a large part of the computational challenge is related to the many equivalent (and hence uninteresting) CRMs. After mining the potential CRMs, several alternative statistical significance tests can be used for ranking in phase 3. As our algorithm imposes no restriction on the size of the CRM, the found CRMs can be quite large for small datasets. It is hence necessary to provide a sufficient number of genes to avoid long CRMs containing false positive TFs.

URL

http://dtai.cs.kuleuven.be/CP4IM/

Presenting Author

Tias Guns (tias.guns@cs.kuleuven.be)
K.U. Leuven

Author Affiliations

(1) DTAI, Department of Computer Science, K.U. Leuven, Belgium. (2) CMPG/BIOI, Department of Microbial and Molecular systems.

Acknowledgements

1) Postdoc and project from the Research Foundation--Flanders, project "Principles of Patternset Mining" 2) PhD and project from the Agency for Innovation by Science and Technology in Flanders (IWT), project "A Bioinformatics Framework for Top-Down Systems Biology" 3) Project IUAP P6/25 (BioMaGNet)

E-55. High-throughput detection of epistasis in studies of the genetics of complex traits

Gyenesei A (1,), Laiho A (1), Semple C (2), Haley C (2), Wei W (2)*

Gene interactions are thought to be important in shaping complex trait variation in agricultural, model organism and human disease genetics. They have been poorly explored, however, because of the lack of high throughput tools to analyze many different traits. The key challenge is how to effectively reduce the search space without losing power of detection so that one trait can be analysed in CPU hours.

Materials and Methods

The developed method is based on a bi-clustering data mining algorithm named Mining Attribute Profiles that has been successfully applied in gene expression data analyses.

Results

We have developed a novel method, called Epicluster, that can quickly select candidate SNPs with consistent genotype distribution patterns differentiating phenotypes of interest and then perform comprehensive epistasis analysis among the candidate SNPs.

Discussion

In the future we plan to implement the developed algorithm as a distributed computing tool to allow multiple traits to be analyzed simultaneously.

Presenting Author

Attila Gyenesei (attila.gyenesei@btk.fi)
Turku Centre for Biotechnology, University of Turku and Åbo Akademi University, Turku, Finland

Author Affiliations

(1) Turku Centre for Biotechnology, University of Turku and Åbo Akademi University, Turku, Finland (2) MRC Human Genetics Unit, Edinburgh, UK

E-56. Heavy metal resistance in Cupriavidus metallidurans: a complex evolutionary and transcriptional process

Monsieurs P (1,), Moors H (1), Van Houdt R (1), Janssen P (1), Mergeay M (1), Leys N (1)*

The soil bacterium Cupriavidus metallidurans CH34 shows great promise in bioremediation schemes since it contains at least 25 loci involved in heavy metal resistance, allowing it to thrive in metal-rich environments. A first step in understanding the molecular mechanisms that underly heavy metal resistance is to reconstruct the transcriptional regulatory networks.

Materials and Methods

We designed a whole-genome microarray to investigate the full stress response of C. metallidurans CH34 at the transcriptomic level when it was challenged to a variety of heavy metals including zinc, copper, cadmium, and lead. By performing a large number of microarray experiments and developing a novel algorithm, additional heavy metal response genes were detected. The redundancy of these genes, which often clustered together within certain genome regions, was assessed by paralogy and phylogenetic analysis.

Results

Certain heavy metal response gene clusters showed similar expression profiles when cells were exposed to varying combinations of heavy metals, thus pointing to complex cross-talk at the transcriptional level between the different heavy metal resistance mechanisms. The highly redundant nature of these heavy metal response gene clusters combined with their phylogenetic distribution within evolutionary related metal resistant bacteria sheds light on the recent evolution of this soil bacterium towards a highly metal-resistant species.

Discussion

The intricate transcriptional behavior leading to metal resistance in C. metallidurans CH34 represents an ideal test case to reconstruct the regulatory networks underlying this response. The complexity of this regulation is underpinned by the complete genome annotation of the C. metallidurans CH34 which allowed us to identify several hundreds of regulatory proteins, many of which are involved in metal detoxification and general metal resistance.Using a combination of microarray results and DNA and protein motifs, we integrated these data to get a better view on the complex response.

Presenting Author

Pieter Monsieurs (pmonsieu@sckcen.be)
Belgian Nuclear Research Center (SCK•CEN)

Author Affiliations

(1) Belgian Nuclear Research Center (SCK-CEN), Belgium.

E-57. Identification of co-regulated mRNA/miRNA pairs by the CoExpress software tool

Nazarov P V (1,), Khutko V, Schmitz S (2), Muller A (1), Kreis S (2), Vallar L (1)*

Small noncoding miRNAs influence most fundamental biological processes by ultimately altering the expression levels of proteins. miRNAs tend to have long half lives and therefore represent promising candidates to be used as disease markers and therapeutic targets. Insights into miRNA functions and miRNA target genes can be obtained from simultaneous analyses of full genome transcription profiles and miRNA levels derived from the same sample types. Therefore, there is a need for effective and user-friendly tools for fast analysis of co-regulation between miRNAs and mRNAs.

Materials and Methods

The co-regulation study was performed using the developed stand-alone software tool CoExpress, which allows for interactive analysis of mRNA and miRNA co-expression by using microarray data. The software performs microarray data pre-processing, building and visualization of co-expression between mRNA and miRNA using correlation or mutual information metrics. Taken together, the software facilitates the analysis and visualization of miRNA-mRNA, mRNA-mRNA and miRNA-miRNA co-expression events, some of which were confirmed by real-time quantitative PCR.

Results

The proper functioning of the software was tested using public mRNA and miRNA expression data from 14 various cell lines. Data from 42 Affymetrix HGU133plus2 arrays and 14 miRNA custom microarray experiments were downloaded from public repositories, normalized and analyzed. We have detected 7423 co-expression events between 2533 mRNAs and 199 miRNAs with r2>0.6. 22 of the most prominent mRNA-miRNA co-expression events were validated by qPCR, which showed good concordance with the results of co-expression analyses.

Discussion

The computational validation of the results was performed by permutation of the samples in the mRNA data set. For the threshold of r2 = 0.6 the estimated p-value < 1e–7. Then, the lists of co-expressed genes were compared to lists of potential miRNA targets, predicted by combination of 6 commonly used algorithms EIMMo, DIANA, Pictar, TargetScan, PITA, and miRanda. Despite the small concordance between predicted targets and co-expressed gene lists, we were able to find negatively regulated genes with significant p-values for most of the 199 considered miRNAs, which showed inverse correlation

URL

http://www.bioinformatics.lu/

Presenting Author

Petr V. Nazarov (petr.nazarov@crp-sante.lu)
Microarray Center, CRP-Santé

Author Affiliations

1. Microarray Center, CRP-Santé, Luxembourg 2. Life Sciences Research Unit, University of Luxembourg

E-58. Identification of genes with preferential expression in the egg cell

*Köszegi D (1, *), Czhial A (1), Kumlehn J (1), Altschmied L (1), Baumlein H (1)*

In contrast to animals, the life cycle of higher plants alternates between a gamete-producing (gametophyte) and a spore-producing generation (sporophyte). The angiosperm female gametophyte consists of four distinct cell types, including two gametes, the egg and the central cell, which give rise to embryo proper and the nutritive endosperm, respectively. To gain insights into the molecular basis of gamete differentiation and function, genes with preferential expression in egg and central cell need to be isolated.

Materials and Methods

A combined subtractive hybridization and virtual subtraction approach was used to isolate egg cell specific genes from a wheat egg cell cDNA library. Using microarray hybridization and in silico subtraction, egg cell expressed genes of Arabidopsis were isolated from a transcription factor induced proliferating tissue, which exhibits an egg cell-like transcriptome.

Results

In total we have isolated seven and nine candidate genes with preferential expression in the wheat and Arabidopsis egg cell, respectively. Via single cell RT-PCR we confirm the preferential expression of three wheat genes. Transgenic Arabidopsis plants transformed with promoter:reporter constructs confirm egg cell specific promoter activity for four out of nine genes.

Discussion

We demonstrate the suitability of the combined subtractive approach for the isolation of gamete specific genes. The approach is broadly applicable also for other species. Isolation and characterization of genes with preferential expression in either gamete allow to unravel the regulatory network which controls specification and differentiation of these important cell types in plants.

Presenting Author

David Köszegi (koszegi@ipk-gatersleben.de)
IPK-Gatersleben

Author Affiliations

1 - IPK-Gatersleben

Acknowledgements

BMBF GABI-SEEDII and InnoRegio Program (No.: 03i0602)

AUTHOR INDEX