Regulatory Genomics Special Interest Group (RegGen SIG)
Program July 20, 2013

07:30		Registration
08:30	5 min	Welcome to SIG
	40 min	Keynote **Stein Aerts**, University of Leuven, Belgium *Motif-based identification of master regulators and direct TF-target interactions in human and* Drosophila *gene networks*
	15 min	**Ivan Kulakovskiy**, Russian Academy of Sciences, Moscow, Russia *diChIPMunk: utilizing ChIP-Seq data to construct advanced dinucleotide models of transcription factor binding sites*
	20 min	**Erik van Nimwegen**, Basel University, Switzerland *The transcription factors democracy: Completely automated inference of genome-wide regulatory interactions from sequencing data*
	20 min	**Alan Moses**, University of Toronto, Canada *Systematic identification of conserved non-coding sequences in plants*
10:15	30 min	Morning Coffee Break
10:45	15 min	**Sunil Kumar**, Swiss Inst. for Exp. Cancer Research, Switzerland *Predicting CTCF site occupancy using sequence and chromatin-associated features*
	20 min	**Struan FA Grant**, University of Pennsylvania, USA *Following functional clues based on the genetic commonalities of diabetes and cancer*
	15 min	**Anirban Bhar**, Georg August University, Göttingen, Germany *Revealing exclusive usage of T-BOX family paralogous transcription factors through identifying diversity in expression profiles during hiPSC-derived cardiomyocytes generation*
	15 min	**Felicia Ng**, Cambridge Institute for Medical Research, UK *Shared transcription factors contribute to stage-specific transcriptional programs during blood cell differentiation*
	20 min	**Andrea Califano**, Columbia University, USA *Assembly and interrogation of tumor-specific regulatory models reveals master regulators of tumor maintenance and chemo-sensitivity*
	15 min	**Anna Lyubetskaya**, Boston University, USA *Reconstructing the regulatory network of TB: Transcription factor binding distribution and properties*
12:30	60 min	Lunch / Poster Session
13:30	40 min	Keynote **Ben Lehner**, Centre for Genomic Regulation, Universitat Pompeu Fabra, Barcelona, Spain *The genetics of individuals: Why should a mutation kill me, but not you?*

	15 min	**Maria Rodriguez Martinez**, Columbia University, USA *GWAS next generation: identifying mechanisms of action in association studies*
	20 min	**Boris Lenhard**, Imperial College London, UK *Alternative and overlapping determinants of transcription start site selection in vertebrate promoters*
	20 min	**Johannes Söding**, Ludwig Maximillian University, Germany Drosophila *Pol II core promoters cluster into four classes characterized by distinct sets of motifs, regulatory properties, and nucleosome patterning*
	20 min	**Jason Ernst**, UCLA, USA *Interplay between chromatin state, regulatory binding, and regulatory motifs*
15:30	30 min	Afternoon Coffee Break / Posters
16:00	20 min	**Bartek Wilczynski**, University of Warsaw, Poland *Predicting regulatory domain boundaries from chromatin immuno-precipitation data*
	15 min	**Guo-Cheng Yuan**, Dana-Farber Cancer Institute, USA *Prediction of chromatin state variability from genomic sequence*
	15 min	**Jan Hapala**, Masaryk University, Brno, Czech Republic *Rotational positioning of regulatory elements within nucleosomes*
	15 min	**Idit Kosti**, Israel Institute of Technology, Haifa, Israel *Does intragenic DNA methylation determine differential exon expression?*
	15 min	**Christoph Kaleta**, University of Jena, Germany *Survival of the quickest – Identification of time-optimal regulatory strategies of metabolism in* Escherichia coli
	15 min	**Lieven Verbeke**, Ghent University, Belgium *EPSILON: localized networks for eQTL prioritization*
	15 min	**Sridhar Hannenhalli**, University of Maryland, USA *Enhancer networks – Hidden layer of gene regulation*
	5 min	Concluding remarks on SIG
18:00		SIG Ends

Oral presentations

Motif-based identification of master regulators and direct TF-target interactions in human and Drosophila gene networks

Stein Aerts

Laboratory of Computational Biology, Department of Human Genetics, University of Leuven, Belgium

We revisit the problem of motif discovery in Metazoan co-expressed gene sets. We discuss in this talk how classical motif discovery, but also modern 'track discovery', can be complementary approaches to ChIP-seq assays and how they continue being invaluable to decipher gene regulatory networks. This is particularly true for biological systems that are less amenable to high-throughput methods, and for processes for which the master regulators are yet unknown. We illustrate the power of motif discovery by mapping an extensive gene regulatory network underlying Drosophila eye development. To this end, we exploit (1) tissue-specific gene expression across three Drosophila species; (2) multiple genetic perturbations and cell sorting experiments in the eye disc; and (3) open chromatin profiling using FAIRE-seq. We identify several new targetomes of eye-related transcription factors, such as Glass, the master regulator of photoreceptor differentiation.

As a next step towards the integration of motif discovery with gene regulatory network inference, we developed iRegulon, a Cytoscape plugin that unites cis-regulatory sequence analysis with biological network tools. Using iRegulon, we re-analyzed microRNA target sets, signaling pathways, Gene Ontology classes, STRING and GeneMania networks, TF perturbation signatures, and finally twenty thousand cancer gene signatures. Through meta-analysis we summarize TF-target interactions yielding "meta-targetomes" that can be useful to annotate re-sequenced cancer genomes.

diChIPMunk: utilizing ChIP-Seq data to construct advanced dinucleotide models of transcription factor binding sites

I. KULAKOVSKIY [1,2,*], V. LEVITSKY [3,4], D. OSCHEPKOV [3], I. VORONTSOV [2,5], V. MAKEEV [1,2,6]

(1) Laboratory of Bioinformatics and Systems Biology, Engelhardt Institute of Molecular Biology, Russian Academy of Sciences, Vavilov str. 32, Moscow, 119991, GSP-1, Russia

(2) Department of Computational Systems Biology, Vavilov Institute of General Genetics, Russian Academy of Sciences, Gubkina str. 3, Moscow, 119991, GSP-1, Russia

(3) Laboratory of Molecular Genetics Systems, Institute of Cytology and Genetics of the Siberian Division of Russian Academy of Sciences, Lavrentiev Prospect 6, Novosibirsk, 630090, Russia

(4) Faculty of Natural Sciences, Novosibirsk State University, Pirogova str. 2, Novosibirsk, 630090, Russia

(5) Yandex Data Analysis School, Data Analysis Department, Moscow Institute of Physics and Technology, Leo Tolstoy Str. 16, Moscow, 119021, Russia

(6) Faculty of Molecular Biology, Moscow Institute of Physics and Technology, Institutskii per. 9, Dolgoprudny, 141700, Moscow Region, Russia

* ivan.kulakovskiy@gmail.com

Abstract

Computational analysis and prediction of transcription factor binding sites (TFBS) is one of the fundamental tasks in regulatory genomics. A TFBS model can be derived from a set of experimentally determined DNA sequences, specifically recognized by a transcription factor (TF). A typical approach is to apply computational *de novo* motif discovery tools. With ChIP-Seq as the new gold standard for genome-wide detection of TFBS *in vivo* it becomes possible to construct advanced TFBS models. Here we present a special motif discovery tool, diChIPMunk, which can produce dinucleotide positional weight matrices (diPWMs) from ChIP-Seq data. We show that diPWMs produced by diChIPMunk significantly outperform existing classic mononucleotide matrices in terms of TFBS recognition quality.

The software is freely available: http://autosome.ru/dichipmunk/

Introduction

Transcription regulation in higher eukaryotes involves transcription factors (TFs) specifically recognizing binding sites (TFBS) in DNA. Experimental techniques based on chromatin immunoprecipitation produce thousands of DNA segments putatively recognized by a TF. One of typical aims is to detect a common text pattern representing preferred TFBS. Careful representation of this pattern, the TFBS model, allows computational prediction of TFBS in genomic sequences of interest.

The most widely used TFBS model is a positional weight matrix (PWM) directly computed from a gapless multiple local alignment of TFBS-containing sequences. PWM assumes independent nucleotide frequencies in different alignment columns, as there were no correlations between them.

At the same time, some more complex models based on ChIP-Seq data provided only incremental improvement over properly trained traditional PWMs [Bi2011].

A matrix of positional weights based on dinucleotide frequencies takes into account correlations of nucleotides in neighboring alignment positions and provides simple extension of the PWM model. Earlier it was already demonstrated that dinucleotide PWMs could outperform classic mononucleotide PWMs if learned from on a reasonably large set of sequences [Levitsky2007]. The remaining step is to properly utilize ChIP-Seq data for model training.

Here we present diChIPMunk, a tool able to produce dinucleotide PWMs based on ChIP-Seq data.

Results

Earlier we presented ChIPMunk [Kulakovskiy2010], an effective algorithm for construction of traditional PWM models based on ChIP-Seq data. ChIPMunk performed efficiently and accurately in several independent benchmarking studies including a recent one of the DREAM consortium [Weirauch2013]. diChIPMunk is based on the same computational engine as ChIPMunk, and thus shares several advantages including usage of ChIP-Seq peak shape (the reads pileup profile) and a support for multi-threaded computations. To utilize ChIPMunk engine diChIPMunk uses a "superalphabet" approach converting initial DNA sequences written in a mononucleotide A-C-G-T alphabet into dinucleotide sequences with a AA-AC-AT…TT alphabet (with each nucleotide included in two neighboring dinucleotides).

To test TFBS recognition quality we have used different ChIP-Seq datasets to compare diChIPMunk models with those of ChIPMunk and PWMs available from public sources.

Here, as a case study, we used top 1000 ChIP-Seq peaks of NANOG and SOX2 TFs published in [Chen2008]. Even ranked peaks were used for model training; odd ranked peaks were used as control true positive sequences. Using a strategy from [Kulakovskiy2013] we have plotted ROC-curves and calculated area-under-curve (AUC) values. Figure 1 presents results of the comparison.

Several other examples of diChIPMunk models evaluation were presented in the corresponding paper [Kulakovskiy2013].

Conclusions

diChIPMunk is able to produce dinucleotide PWMs that perform significantly better than mononucleotide PWMs. We provide diChIPMunk as a production-ready tool. diChIPMunk is going to be included in BioUML platform [Kolpakov2006] as a motif discovery algorithm along with several accompanying tools. As the dinucleotide PWM is a fairly simple model it becomes possible to adapt many existing supporting tools, such as TFBS prediction in a given sequence (i.e. motif finding), computing P-values for given score threshold levels etc for dinucleotide PWMs. We believe this will facilitate a wider usage of dinucleotide PWMs with more and moreChIP-Seq data becoming available.

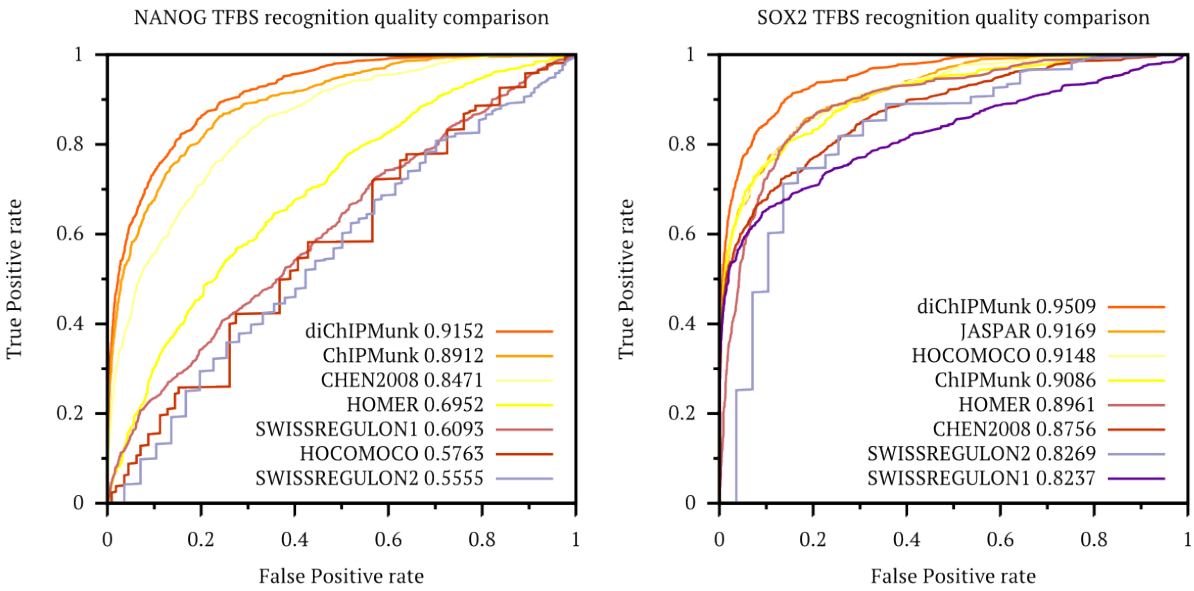


Figure 1. ROC curves of TFBS models for NANOG (left panel) and SOX2 (right panel) TFs. True positive rate was estimated using independent control subset of ChIP-Seq peaks. False positive rate was estimated based on PWM/dinucleotide PWM P-values as described in [Kulakovskiy2013]. AUC values are given in figure legends. HOMER, SwissRegulon and JASPAR PWMs were taken from corresponding collections. CHEN2008 PWM was presented in the same paper as the TF ChIP-Seq data. The SOX2 matrix from JASPAR collection was based on the same ChIP-Seq dataset.

Acknowledgements

This work was supported by a Dynasty Foundation Fellowship [to I.K.]; Russian Foundation for Basic Research [12-04-32082-mol_a to I.K.] and [12-04-01736-a to D.O.]; Presidium of the Russian Academy of Sciences program in Cellular and Molecular Biology.

References

[Bi2011] PLoS One. 2011;6(9):e24210. doi: 10.1371/journal.pone.0024210. Epub 2011 Sep 2. Tree-based position weight matrix approach to model transcription factor binding site profiles. Bi Y, Kim H, Gupta R, Davuluri RV.

[Chen2008] Cell. 2008 Jun 13;133(6):1106-17. doi: 10.1016/j.cell.2008.04.043. Integration of external signaling pathways with the core transcriptional network in embryonic stem cells. Chen X, Xu H, Yuan P, Fang F, Huss M, Vega VB, Wong E, Orlov YL, Zhang W, Jiang J, Loh YH, Yeo HC, Yeo ZX, Narang V, Govindarajan KR, Leong B, Shahab A, Ruan Y, Bourque G, Sung WK, Clarke ND, Wei CL, Ng HH.

[Kolpakov2006] Proceedings of The Fifth International Conference on Bioinformatics of Genome Regulation and Structure; July 16–22, 2006; Novosibirsk, Russia. 2006.3 p. 281-285. BioUML: visual modeling, automated code generation and simulation of biological systems. Kolpakov F, Puzanov M, Koshukov A.

[Kulakovskiy2010] Bioinformatics. 2010 Oct 15;26(20):2622-3. doi: 10.1093/bioinformatics/btq488. Epub 2010 Aug 24. Deep and wide digging for binding motifs in ChIP-Seq data. Kulakovskiy IV, Boeva VA, Favorov AV, Makeev VJ.

[Kulakovskiy2013] J Bioinform Comput Biol. 2013 Feb;11(1):1340004. doi: 10.1142/S0219720013400040. Epub 2013 Jan 16. From binding motifs in ChIP-Seq data to improved models of transcription factor binding sites. Kulakovskiy I, Levitsky V, Oshchepkov D, Bryzgalov L, Vorontsov I, Makeev V.

[Levitsky2007] BMC Bioinformatics. 2007 Dec 19;8:481. Effective transcription factor binding site prediction using a combination of optimization, a genetic algorithm and discriminant analysis to capture distant interactions. Levitsky VG, Ignatieva EV, Ananko EA, Turnaev II, Merkulova TI, Kolchanov NA, Hodgman TC.

[Weirauch2013] Nat Biotechnol. 2013 Feb;31(2):126-34. doi: 10.1038/nbt.2486. Epub 2013 Jan 27. Evaluation of methods for modeling transcription factor sequence specificity. Weirauch MT, Cote A, Norel R, Annala M, Zhao Y, Riley TR, Saez-Rodriguez J, Cokelaer T, Vedenko A, Talukder S; DREAM5 Consortium, Bussemaker HJ, Morris QD, Bulyk ML, Stolovitzky G, Hughes TR.

The transcription factors democracy: Completely automated inference of genome-wide regulatory interactions from sequencing data

Erik van Nimwegen

Center for Molecular Life Sciences, Basel University, Switzerland

How do gene regulatory networks control cell fate and identity in higher eukaryotic organisms? Although gene expression and chromatin state dynamics are ultimately encoded by constellations of binding sites recognized by regulators such as transcriptions factors (TFs) and microRNAs (miRNAs), our understanding of this regulatory code and its context-dependent read-out remains very limited. Experimental researchers interested in elucidating the key regulatory interactions acting within a particular biological system of interest face the difficulty that in higher eukaryotes, there are thousands of potential regulators, and it is not feasible to investigate all these using direct experimentation. Although it has become relatively straight-forward, using next-generation sequencing, to obtain genome-wide measurements of gene expression, chromatin state, and TF-binding dynamics, it is typically far beyond the expertise of experimental groups to connect such data to the actions of individual regulators. And even when experimentalists team up with expert computational biologists, inferring key regulators and their genome-wide interactions from high-throughput data remains highly challenging, typically involving `case-by-case' development of methodology.

In recent years we have developed a methodology that combines automated processing of next-generation sequencing data with genome-wide predictions of TF binding sites and miRNA target sites to model gene expression or chromatin modifications in terms of these sites. This completely automated system, called ISMARA, is available through a web-interface (ismara.unibas.ch) and requires only the uploading of raw micro-array or sequencing (RNA-seq or ChIP-seq) data. ISMARA then automatically identifies the key TFs and miRNAs driving expression/chromatin changes and makes detailed predictions regarding their regulatory roles. These include predicted activities of the regulators across the samples, their genome-wide targets, enriched gene categories among the targets, and direct interactions between the regulators.

In the presentation I will discuss various aspects of the methodology implemented in ISMARA, and illustrate the power of the approach by demonstrating that, for well-studied model systems, ISMARA consistently identifies known key regulators and their actions ab initio in a completely automated fashion and without any tunable parameters.

Systematic identification of conserved non-coding sequences in plants

Alan Moses

Department of Cell and Systems Biology, University of Toronto, Canada

Despite the central importance of noncoding DNA in gene regulation and evolution, our understanding of the genomic extent and nature of selection on plant noncoding regions remains limited. This is in contrast to other clades containing model organisms (mammals, fruit fly, budding yeast, etc.) where studies of sequence conservation across large numbers of related genomes have provided a powerful approach to identify and characterize functional noncoding sequences. To systematically identify conserved non-coding regions in Arabidopsis and its close relatives (crucifers) we sequenced three Brassicaceae species and analyzed them alongside six previously sequenced crucifer genomes. We compared the conservation of non-coding DNA in plants to what had been previously observed in other organisms. For example, although we find that these plants have shorter and fewer conserved non-coding sequences than have been observed in animals, genes involved in development, particularly transcription factors, are associated with large numbers of the most highly constrained non-coding sequences. Remarkably, since plants' and animals' most recent common ancestor was likely unicellular, this suggests that complex regulatory control of developmental patterning transcription factors evolved independently in the two major lineages of complex multicellular life. We also performed whole genome motif-finding on the conserved non-coding sequences and identified known and novel transcription factor binding specificities, as well as other motifs. Finally, using population genomics data, we tested for more recent evidence of selection on the conserved non-coding sequences and regulatory motifs.

Predicting CTCF site occupancy using sequence and chromatin-associated features

Sunil Kumar[1,2], René Dreos[2], Giovanna Ambrosini[1,2], Philipp Bucher[1,2]
[1]*Swiss Institute for Experimental Cancer Research (ISREC), School of Life Sciences, EPFL, Station 15, Lausanne CH-1015, Switzerland*
[2]*Swiss Institute of Bioinformatics (SIB), EPFL, Station 15, Lausanne CH-1015, Switzerland*

Introduction

Genes are regulated by transcription factors (TF) binding to physiological target sites in the genome. Understanding the mechanisms by which TFs are recruited to their target sites is essential for the understanding of gene regulation. The recently introduced ChIP-Seq technology allows for genome-wide mapping of all *in vivo* bound sites of a given TFs in a particular cell type at near base-pair resolution [1]. What has become clear from ChIP-Seq experiments is that the intrinsic binding specificity of a TF can only partly explain the *in vivo* site occupancy patterns, which were found to be remarkably tissue-specific.

Several recent studies have reported that TF binding is influenced and can be chromatin contextual features such as DNA chromatin accessibility, nucleosome occupancy, or the presence of specific histone post-translational modifications [2,3]. Site occupancy may also be partly predictable from sequence intrinsic properties such as oligo-nucleotide composition, DNA structural parameters and evolutionary conservation. In this work, we use machine learning to assess the relative importance of such features in TF to target site recruitment process, in the hope to gain insights into transcription regulatory mechanisms. As test example, we use the sequence-specific DNA-binding protein CTCF which has been assayed by ChIP-Seq in many cell types. CTCF has been attributed diverse roles in gene regulation, including insulator activity, gene activation and repression, genomic imprinting and tumor suppressor [4].

Study design, data and methods

Our study primarily relies on recent data published by the ENCODE consortium [5]. The general idea is to use machine learning algorithms to build models that predict site occupancy at predefined target sites. We used two types of candidate target site list: (a) Predicted sites from a whole genome scan with a position weight matrix (PWM) and (b) 10 cell type-specific peak lists published by ENCODE. For both types of candidate sites we expressed cell-type specific CTCF occupancy as the number of ChIP-Seq tags within a window of 200 bp around the site. In parallel, we collected for each site in each lists a number of associated predicted and experimental features (Table 1). We then applied machine learning algorithms to predict site occupancy from different subsets of associated features. Note that experimental features such as histone modifications or DNase I hypersensitivity were evaluated in two different ways: (a) by the total number of tag counts in a window around the site and (b) by computing a "shape-score" reflecting the similarity of the tag distribution around a particular site with the average tag distribution as seen in an aggregate plot. We applied machine learning in a binary class-prediction framework and by regression analysis. For class prediction the candidate site lists were first split into high and low occupancy classes. Support vector machines (SVM) combined with recursive feature selection performed better than random forests (RF) was used for class prediction, support vector regression (SVR) was used for model training with quantitative site occupancy data. The performance was evaluated by 10-fold cross-validation, except in the cases where data from one cell type were used to predict the results from another cell type. Performance was expressed as a Pearson correlation coefficient between predicted and experimentally determined site occupancy (Fig. 1A).

Summary of results and selected examples

Rad21, TFBS-score and DGF are the most important features that contributed significantly to the classification followed by histone marks. It was known before that CTCF associates with Rad21 in the so-called CTCF/cohesion complex. The very good performance of Rad21 confirms the previous reports that CTCF acts in close coordination with this protein. Other sequence and structural features showed relatively low importance and didn't contribute significantly to the classification accuracy when considering TBFS-score alone, however, they performed better in predicted CTCF sites than ENCODE peak list. In order to get an objective comparison between the two datasets we used top two features (showing maximum contribution) to build model on one cell line and prediction on other cell lines. In addition, five histone marks were also used for classification in five cell lines (limited due to dataset availability), where H3K4me1 a distal mark showed highest importance in classification. The two cancerous cell lines in the dataset K562 and HepG2 showed similar patterns among themselves but a distinct pattern when compared to other normal cell lines indicating that they has varied cell specific CTCF sites.

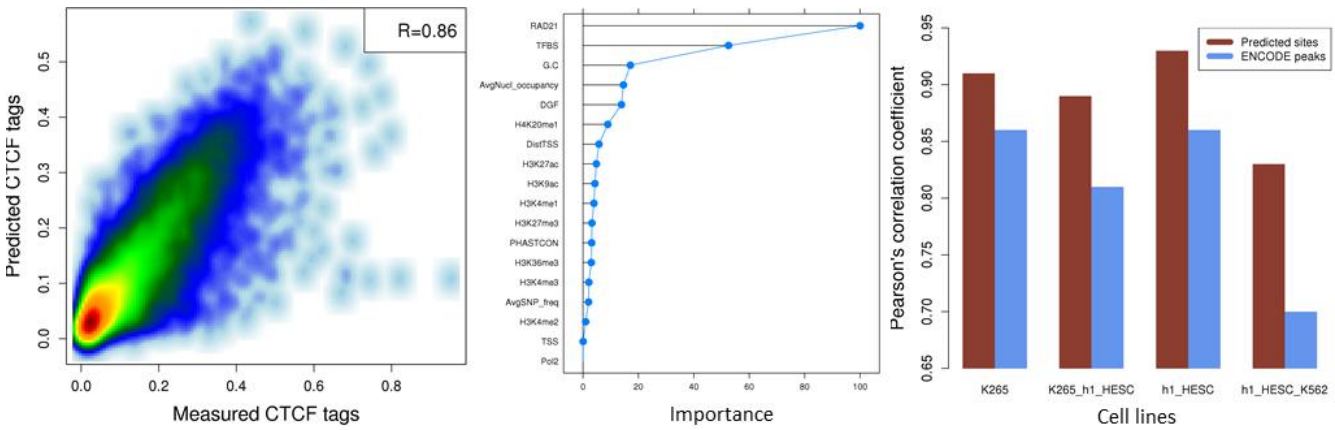


Figure1: A) SVR model was built on training dataset from K562 and used to predict CTCF tags on test dataset from K562 cell line. Graph shows the Pearson's correlation for measured and predicted CTCF sites on test dataset. B) Feature importance for regression on CTCF sites from ENCODE peak list for K562 cell line. C) SVR model was built on K562 and H1-hESC cell line for both dataset (ENCODE peak list and predicted CTCF sites), these models were then used to predict the CTCF tags on other cell line (Pearson's correlation coefficient shows the performance of model on cross-cell line prediction).

Table 1: Overview of features used to predict CTCF site occupancy

Feature type	**Feature description**	**Number**	**Data source/ Reference**
Sequence intrinsic	Mono-penta nucleotide frequency	1364	
	Nucleosome occupancy	1	[6]
	Structural parameters	10	[7]
	CTCF TFBS-score	1	JASPAR [8]
Evolution/Polymorphism	Avg PhastCons score	1	UCSC database
	SNP Frequency	1	dbSNP132
Tissue-specific experimental features (tag counts in window around site and shape based evaluation)	Distance to nearest TSS	1	ENSEMBL database
	DNase I	1	GEO series (GSE26328)
	Histone modifications	8	GEO series (GSE29611)
	PolII	1	GEO series (GSE32465)
	RAD21	1	GEO series (GSE32465)

To investigate the degree of tissue specificity we built models using TFBS-score, DGF, Rad21, average nucleosome occupancy/ base, average distance from TSS site and certain histone marks performed best in the prediction (Figure 1B). The results showed in general high correlation coefficient within and

across cell lines with certain notable exception (Figure 1C). Models built from fibroblast cell lines were good predictor of other fibroblast cell lines compared to other cell lines. The other cell line which showed consistently varied performance in cross-prediction experiment is K562, the cancer cell line considered in this study. We also note that tissue-specificity of trained models are better predictor of PWM predicted sites lists than experimental peak lists. This is not really surprising as the experimental peak list exclude those sites which totally unoccupied in given tissue.

References

[1] Valouev et al., 2008, Nat Methods, 5, 829-834
[2] Barski et al., 2007, Cell, 129, 823-836
[3] Neph at al., 2012, Nature, 489, 83-90
[4] Lee et al., 2012, J Biol Chem, 287, 30906-30913
[5] The ENCODE Project Consortium, 2012, Nature, 489, 57-74
[6] Kaplan et al., 2009, Nature, 458, 362-366
[7] Florquin et al., 2005, Nucleic Acid Res, 33, 4255-4264
[8] Portales-Casamar et al. 2010, Nucleic Acids Res, 38, D105–D110

Following functional clues based on the genetic commonalities of diabetes and cancer

Struan F.A. Grant

Division of Human Genetics, Children's Hospital of Philadelphia Research Institute, Perelman School of Medicine, University of Pennsylvania, USA

The repertoire of genes already established to play a role in the pathogenesis of type 2 diabetes (T2D) has grown substantially due to recent genome wide association studies (GWAS). In 2006, we discovered the strong association of variants in the transcription factor 7 like 2 (*TCF7L2*) gene with T2D. Other investigators have already independently replicated this finding in different ethnicities and, interestingly, from the first GWAS of T2D in Caucasians, the strongest association was indeed with *TCF7L2*; this is now considered the most significant genetic finding in T2D to date.

Interestingly, there is also a very strong connection between *TCF7L2* and cancer. The key 8q24 locus found to be the most strongly associated genomic region with a number of cancers through GWAS contributes to the disease pathogenesis through mutation of an upstream TCF7L2-binding element driving the transcription of the *MYC* gene. Indeed, is has been known for some years that *TCF7L2* harbors specific mutations that strongly influence colorectal cancer risk plus genomic sequencing of colorectal adenocarcinomas identified a recurrent *VTI1A-TCF7L2* gene fusion. Furthermore, many of the T2D GWAS-derived risk conferring alleles have been shown to protect against prostate cancer; in addition, *THADA*, *JAZF1* and *TCF2* are loci that have been strongly detected in separate GWAS analyses of prostate cancer and T2D. Thus, *TCF7L2* and other T2D associated genes also appear to be key players in cancer pathogenesis; however, this mechanism is still far from understood.

We previously performed chromatin immunoprecipitation followed by high-throughput sequencing (ChIP-seq) with this transcription factor to elucidate its binding repertoire genome wide. Unexpectedly, and despite employing a carcinoma cell line, the genes with TCF7L2 binding sites are strongly enriched in pathway categories related to metabolic-related functions and traits, further suggesting a role for metabolism in cancer. Furthermore, the list of loci bound by TCF7L2 harbors a highly significant over-representation of GWAS loci associated with T2D and cardiovascular disease.

With all these intriguing facts in mind, we are taking forward the loci that are common to T2D and cancer GWAS outcomes and investigating the impact on cell proliferation with the ultimate goal of testing their role in beta-cell proliferation in mice, a mechanism which still largely eludes the diabetes research community.

Revealing Exclusive Usage Of T-BOX Family Paralogous Transcription Factors Through Identifying Diversity In Expression Profiles During hiPSC-Derived Cardiomyocytes Generation

Anirban Bhar
Institute of Bioinformatics, University Medical Center Goettingen, Georg August University, Goettingen, Germany
anirban.bhar@bioinf.med.uni-goettingen.de

Martin Haubrock
Institute of Bioinformatics, University Medical Center Goettingen, Georg August University, Goettingen, Germany
martin.haubrock@bioinf.med.uni-goettingen.de

Edgar Wingender *
Institute of Bioinformatics, University Medical Center Goettingen, Georg August University, Goettingen, Germany
edgar.wingender@bioinf.med.uni-goettingen.de

MOTIVATION

Functional genomics aims to understand dynamic features encoded in the genome such as transcription of genes, thereby frequently using results from high throughput approaches. Transcription, RNA splicing and translation are the key steps in the process of gene expression. Production of a specific gene product can be increased or decreased by regulation of any of these steps. DNA microarrays are used to measure expression levels of a large number of genes simultaneously over a set of experimental conditions. In recent years, expression levels of thousands of genes are not only measured over sets of experimental conditions but also across many time points. To analyze such high throughput 3D datasets we need computational approaches. Coexpression analysis helps to retrieve functionally coherent group of genes that are often coregulated by a common transcription factor. Clustering, one of the unsupervised learning approaches can retrieve a group of genes having similar expression profiles over all experimental conditions. But it has been observed that genes are not necessarily to be coexpressed over all samples in a gene expression dataset, i.e.- genes can have similar expression profiles over a subset of samples. To simultaneously group genes and samples, biclustering or subspace clustering methods are used. However, biclustering algorithms fail to cluster genes, samples and time points simultaneously in a time series gene expression data. To cope with that problem triclustering algorithms are used. Zhao et al. proposed a triclustering algorithm TRICLUSTER to find groups of coexpressed genes in such time-series gene expression data set [1]. Tchagang et. al. recently proposed OPTricluster algorithm that is also able to cluster genes, samples and time points simultaneously [2]. One of the limitations of OPTricluster is that it can only cope with short time series gene expression datasets. In our previous work we have proposed triclustering algorithm δ-TRIMAX to mine such 3D gene expression datasets by introducing a novel definition of mean squared residue score for mining 3D datasets [3]. The goal of δ-TRIMAX is to retrieve maximal triclusters having mean squared residue score below a threshold δ. The limitations of δ-TRIMAX is that it is unable to extract overlapping triclusters. As δ-TRIMAX replaces each element of tricluster found in one iteration by random numbers, it can affect the originality of the dataset. In this paper we introduce the triclustering algorithm EMOA-δ-TRIMAX that can retrieve a group of genes that are coexpressed and coregulated over a subset of samples across a subset of time points. Here we have used Non-dominated Sorting Genetic Algorithm-II (NSGA-II) to balance the trade-off between the aforementioned conflicting objectives i.e. minimizing mean squared residue score, maximizing volume of the triclusters and generate pareto optimal solutions that are equally distributed in the objective space [4]. Additionally we have also maximized Spearman correlation coefficient of resultant triclusters. Our proposed algorithm also effectively deals with the drawbacks of our previously proposed algorithm δ-TRIMAX.

Regulation of transcription by transcription factors (TFs) can be initiated through binding to defined cis-regulatory elements in promoters. For accomplishing the function as an activator or inhibitor, TFs must recognize the regions where they should bind to and they do so through DNA-binding domains (DBD) [5]. A systematic classification of TFs according to their DBDs can help to predict the DNA-binding specificity of TFs with as yet ill-characterized DNA-binding properties. Paralogous transcription factors may have derived from a common ancestor by a gene duplication event and these transcription factors are assumed to participate in a novel function or some specialized ones of their original functions. Many of them still share major properties of their DBD and, thus, bind to identical or highly related cis-regulatory elements [5]. Mutation of the activation domain of paralogous transcription factors may yield alteration of their interacting partners in spite of having similar DNA-binding domains. Divergence of

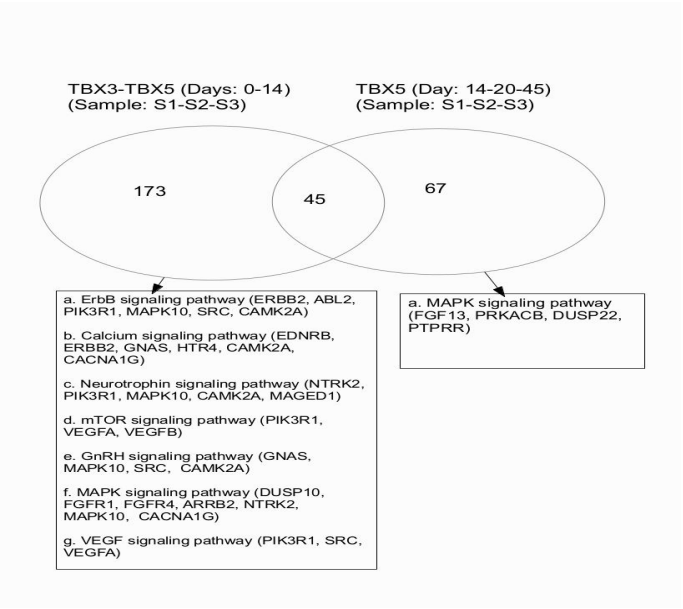


Figure 1: Differentially expressed targets of paralogous transcription factor across different subset of time points

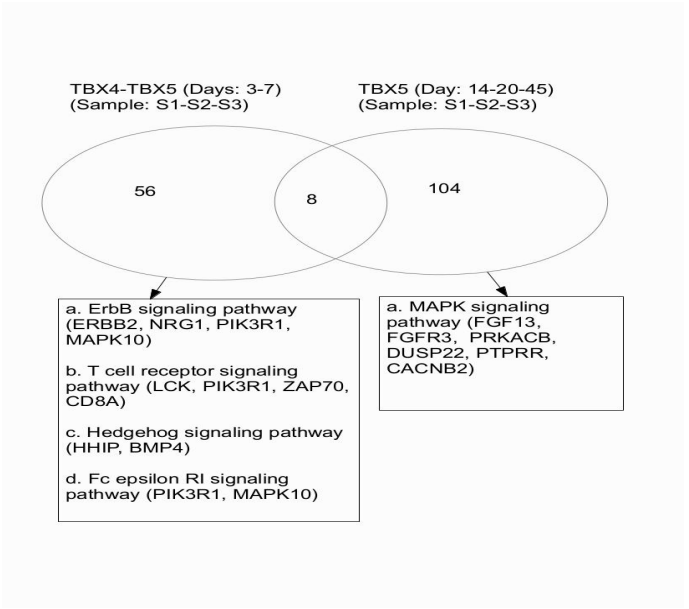


Figure 2: Differentially expressed targets of paralogous transcription factor across different subset of time points

expression profiles of paralogous transcription factors across tissues or time points can be a cause for participating in distinct pathways or regulating the same genes across different tissues or time points. For instance it has been previously reported that two paralogous transcription factors *Pax2* and *Pax3* regulate the gene c-Ret in kidney and neural crest, respectively [6]. Though recent works reveal roles of cardiac transcription factors in molecular regulation of pluripotent stem cell derived cardiomyocytes differentiation, the roles of cardiac paralogous T-Box family transcription factors are still poorly understood during different stages of cardiac differentiation.

RESULT

In this work we have applied our proposed EMOA-δ-TRIMAX algorithm on a time series gene expression dataset that contains mRNA expression profiles during differentiation of human induced pluripotent stem cell (hiPSC) derived cardiomyocytes. This dataset contains 48803 Illumina probe ids, 12 time points (day 0, 3, 7, 10, 14, 20, 28, 35, 45, 60, 90, 120) and 3 samples (GEO accession number GSE35671). Expression values at each time point were generated by three independent runs (Run 1-3) [7]. Our algorithm results in 100 triclusters that cover 88.14% of all probe-ids, 100% of all time points and 100% of all samples. We could show that EMOA-δ-TRIMAX outperforms other triclustering algorithms. It has been reported in the original work that the differentiation of hiPSCs to cardiomyocytes was observed during days 0, 3, 7, 10, 14, 20, 28 and on day 14 heart beating was first perceived. Days 35, 45, 60, 90 and 120 are reported as post-differentiation time points [7]. To establish biological significance of group of co-expressed genes, we checked for KEGG pathway and transcription factor binding site (TFBS) enrichment, the latter by using the TRANSFAC library (version 2012.2) [8]. We used an internal database of around 52 million TFBS predictions that have high affinity scores and are conserved between human, mouse, dog and cow [9]. Out of these 52 million conserved TFBSs we have selected the best 1% for each TRANSFAC matrix individually to select the most specific regulator (transcription factor) - target interactions. We have observed KEGG pathway and TFBS enrichment for 100% and 98% of resultant triclusters, respectively. Through our analysis we identified similar expression profiles of paralogous TFs *TBX3* and *TBX5* across days 0, 14 but divergence in their expression profiles across days 14, 20, 45 over all samples. Figure 1 shows that at early time points both *TBX3*, *TBX5* and at later time points only *TBX5* regulate target genes that participate in distinct sets of pathways. Additionally we observed that both *TBX3*, *TBX5* and only *TBX5* regulate MAPK signaling pathways through binding promoter regions of different target genes at early and later time points, respectively. We also observed similar expression profiles of paralogous transcription factors *TBX4* and *TBX5* across days 3, 7 but divergence in their expression profiles across days 14, 20, 45 over all samples. In Figure 2 we can observe that at early time points both *TBX4*, *TBX5* and at later time point only *TBX5* regulate distinct sets of genes that participate almost different signaling pathways. It has been reported in previous studies that ErbB, calcium, neurotrophin, VEGF, hedgehog signaling pathways play critical roles in cardiac differentiation and development [10–14]. It has been revealed in a previous study that *TBX5* plays a crucial role in

embryonic cardiac cell cycle progression and depletion of *TBX5* leads to cardiac programmed cell death [15]. Interestingly through our analysis we also observed that *TBX5* is expressed in both early and later time points.

CONCLUSION

Our integrated systems biology approach reveals exclusive usage of paralogous transcription factors of the T-BOX family through identifying diversity of their expression profiles and provides new insights into their roles in regulating cardiac differentiation.

REFERENCES

1. L. Zhao and M. J. Zaki, "Tricluster: an effective algorithm for mining coherent clusters in 3d microarray data," in *Proceedings of the 2005 ACM SIGMOD international conference on Management of data*, pp. 694–705, 2005. ISBN:1-59593-060-4.

2. A. Tchagang, S. Phan, F. Famili, H. Shearer, P. Fobert, Y. Huang, J. Zou, D. Huang, A. Cutler, Z. Liu, and Y. Pan, "Mining biological information from 3d short time-series gene expression data: the optricluster algorithm," *BMC Bioinformatics*, vol. 13, April 4 2012.

3. A. Bhar, M. Haubrock, A. Mukhopadhyay, U. Maulik, S. Bandyopadhyay, and E. Wingender, "Coexpression and coregulation analysis of time-series gene expression data in estrogen-induced breast cancer cell," *Algorithms for Molecular Biology*, vol. 8, March 23 2013.

4. K. Deb, A. Pratap, S. Agarwal, and T. Meyarivan, "A fast and elitist multiobjective genetic algorithm: Nsga-ii," *IEEE transactions on evolutionary computation*, vol. 6, no. 2, pp. 182–197, 2002.

5. E. Wingender, T. Schoeps, and J. Dnitz, "Tfclass: an expandable hierarchical classification of human transcription factors," *Nucleic Acids Research*, vol. 41, no. D1, pp. D165–D170, 2013.

6. L. Singh and S. Hannenhalli, "Functional diversification of paralogous transcription factors via divergence in dna binding site motif and in expression," *PloSOne*, vol. 3, no. 6, 2008.

7. B. JE, R. M, S. S, R. P, S. B, B. H, W. T, C. E, C. U, and K. KL, "Determination of the human cardiomyocyte mrna and mirna differentiation network by fine-scale profiling," *Stem Cell Development*, vol. 21, pp. 1956–1965, July 20 2012.

8. E. Wingender, X. Chen, E. Fricke, R. Geffers, R. Hehl, I. Liebich, M. Krull, V. Matys, H. Michael, R. Ohnhuser, M. Prss, F. Schacherer, S. Thiele, , and S. Urbach, "The transfac system on gene expression regulation," *Nucleic Acids Research*, vol. 29, pp. 281–283, January 1 2001.

9. X. Xie, J. Lu, E. Kulbokas, T. Golub, V. Mootha, K. Lindblad-Toh, E. Lander, , and M. Kellis, "Systematic discovery of regulatory motifs in human promoters and 3 utrs by comparison of several mammals," *Nature*, vol. 434, no. 7031, pp. 338–345, 2005.

10. R. LF, "Neurotrophin-regulated signalling pathways," *Philosophical Transactions of the Royal Society B: Biological Sciences*, vol. 361, pp. 1545–1564, September 29 2006.

11. M. Bekhite, A. Finkensieper, S. Binas, J. Mller, R. Wetzker, H. Figulla, H. Sauer, and M. Wartenberg, "Vegf-mediated pi3k class ia and pkc signaling in cardiomyogenesis and vasculogenesis of mouse embryonic stem cells," *Journal of Cell Science*, vol. 124, pp. 1819–1830, June 1 2011.

12. J. Fu, H. Yu, R. Wang, J. Liang, and H. Yang, "Developmental regulation of intracellular calcium transients during cardiomyocyte differentiation of mouse embryonic stem cells," *Acta Pharmacologica Sinica*, vol. 27, pp. 901–910, 2006.

13. F. Rochais, K. Mesbah, and R. Kelly, "Signaling pathways controlling second heart field development," *Circulation Research*, vol. 104, pp. 933–942, 2009.

14. W. Zhu, Y. Xie, K. Moyes, J. Gold, B. Askari, and M. Laflamme, "Neuregulin/erbb signaling regulates cardiac subtype specification in differentiating human embryonic stem cells," *Circulation Research*, vol. 107, pp. 776–786, September 17 2010.

15. S. Goetz, D. Brown, and F. Conlon, "Tbx5 is required for embryonic cardiac cell cycle progression," *Development*, vol. 133, pp. 2575–2584, July 2006.

Shared Transcription Factors Contribute to Stage-specific Transcriptional Programs during Blood Cell Differentiation

Felicia Ng, Fernando Calero-Nieto, Nicola Wilson, Rebecca Hannah, Evangelia Diamanti, Berthold Gottgens

Cambridge Institute for Medical Research, Wellcome Trust/MRC Building, Addenbrooke's Hospital, Hills Road, Cambridge, CB2 0XY United Kingdom

BACKGROUND

Transcription factors (TFs) have long been recognized as important regulators of haematopoietic cell type identity. Specific TFs have been shown to be critical for regulating pluripotency genes in haematopoietic stem cells while others drive differentiation to mature haematopoietic cell types (Orkin and Zon 2008). As a result, TFs have been extensively studied at all stages of haematopoietic development (Wilson et al. 2011). Furthermore, advances in the generation of TF binding maps by ChIP-seq permitted investigations at the genome level. While an abundance of ChIP-seq data exists for different haematopoietic cell types, not much is known about the genome-wide impact of TF binding in driving transcriptional programs of multiple cell types.

An observation from several independent ChIP-seq studies is the strong cell-type-specific binding pattern displayed by many haematopoietic TFs (Hannah et al. 2011; Wei et al. 2011; Pilon et al. 2011). These studies demonstrated that the binding profile of different TFs in the same cell type show stronger correlation than the binding profiles of the same TFs in different cell types. Interestingly, this cell-type-specific binding pattern was also observed for so-called 'master regulators' of haematopoietic stem cells, therefore, raising the question as to how 'master regulators' dictate cell type identity? Most importantly, is this observation an indication of 'functional' rather than 'opportunistic' binding events?

To address these questions, we have analysed the genome-wide binding maps of 10 key haematopoietic stem cell TFs in both primary mast cultures and a progenitor cell line. In addition, expression profiling by RNA-seq on both cell types were analysed in conjunction with the TF binding data to provide a more comprehensive view of gene expression regulation.

RESULTS

Gene expression profiling in mast cells and a progenitor cell line (HPC7) showed that many haematopoietic stem cells 'master regulators' were indeed expressed at similar levels in both cell

types. We also showed that HPC7 closely resembles common myeloid progenitors (precursors of mast cells) and recapitulates the gene expression profile of early blood stem/progenitor cells. Shared expression of key stem cell TFs, therefore, suggests that a more detailed comparative analysis of genome-wide binding patterns in both cell types may provide new insights into the transcriptional control of cell type identity.

A global comparison of HPC7 and mast ChIP-seq data for 10 stem cell TFs (Ctcf, E2a, Erg, Fli1, Gata2, Lmo2, Meis1, PU.1, Runx1, Scl) revealed very little overlap in binding sites (<30%). Moreover, pairwise correlation analysis of all 20 genome wide binding profiles followed by hierarchical clustering revealed clustering of all TFs by cell type, with the exception of Ctcf. These observations suggest that binding of the shared TFs are largely cell-type-specific for 2 closely related haematopoietic cell types. Having identified predominantly cell-type-specific binding patterns for key regulatory TFs raised the question as to whether TFs are passively recruited to cell-type-specific genomic regions of open chromatin with no major regulatory impact or actively participate in 2 different transcriptional programmes. To evaluate the extent to which cell-type-specific binding of shared TFs might be associated with gene expression, we developed multivariate linear regression models to correlate changes in TF binding (ΔTF) with changes in gene expression (ΔGE). Fitting in a simple linear regression model showed some correlation between ΔTF and ΔGE (R2 value ~22.7%). Further application of the linear model on subsets of the data – genes with at least 5TFs bound – increased the R2 value up to ~41.4%. Although higher variability was explained, this is not ideal since many genes were thrown out. We then sought an alternative approach by using generalized additive models (GAM) and by incorporating all pairwise interaction of shared TFs to account for cooperation between TFs. This approach allowed us to fit concordant pairs of TFs to differential gene expression in a non-linear fashion. GAM with interaction terms correlated more strongly with gene expression changes (R^2 ~ 41.8%) than GAM without interaction terms (R2 ~ 25.4%). We were also able to identify interesting TF pairs that co-operate to affect cell-type-specific gene expression.

The modelling approach suggested that cell type specific binding of shared TFs makes meaningful contributions to differential gene expression. However, it remained unclear whether cell-type-specific binding is largely mediated through direct or indirect binding to DNA. To do this, we carried out a comprehensive motif analysis of common as well as cell-type-specific TF-bound regions. We found that consensus sequence motifs of shared TFs were enriched across common and cell-type-specific regions indicating direct DNA binding of the shared TFs. Does this then suggest that cell-type-specific TFs are driving reorganization of shared TFs to cell-type-specific sites? Indeed, we observed specific enrichment and depletion of motifs in cell-type specific regions. From this

analysis, Mitf and c-Fos emerged as potential candidate regulators because their motifs were enriched only in mast-specific regions and our RNA-seq data showed significant over-expression of these genes in mast cells. We went on to generate ChIP-seq data for these 2 mast-specific factors and analysed overlapping binding sites with the 10 shared TFs in HPC7 and mast cells. We were able to show that Mitf and c-Fos binding co-occupy a substantial proportion of regions bound by shared TFs in both cell types but not HPC7-specific regions. Mitf and c-Fos also bind to mast-specific regions, and this 'new' binding is accompanied by relocation of shared TFs to these regions.

CONCLUSION

Taken together, these data are consistent with a model whereby mast cell specific and shared TFs contribute to gene regulation in mast cells by binding to both common and mast cell specific regulatory regions. A comprehensive understanding of how TFs interact with the genome will not only advance basic research but improves our mechanistic understanding of cellular reprogramming strategies developed within the stem cell and regenerative medicine area.

REFERENCES

Hannah R, Joshi A, Wilson NK, Kinston S, Göttgens B. 2011. A compendium of genome-wide hematopoietic transcription factor maps supports the identification of gene regulatory control mechanisms. *Experimental hematology* **39**: 531–41.

Orkin SH, Zon LI. 2008. Hematopoiesis: an evolving paradigm for stem cell biology. *Cell* **132**: 631–44.

Pilon AM, Ajay SS, Kumar SA, Steiner LA, Cherukuri PF, Wincovitch S, Anderson SM, Mullikin JC, Gallagher PG, Hardison RC, et al. 2011. Genome-wide ChIP-Seq reveals a dramatic shift in the binding of the transcription factor erythroid Kruppel-like factor during erythrocyte differentiation. *Blood* **118**: e139–48.

Wei G, Abraham BJ, Yagi R, Jothi R, Cui K, Sharma S, Narlikar L, Northrup DL, Tang Q, Paul WE, et al. 2011. Genome-wide analyses of transcription factor GATA3-mediated gene regulation in distinct T cell types. *Immunity* **35**: 299–311.

Wilson NK, Tijssen MR, Göttgens B. 2011. Deciphering transcriptional control mechanisms in hematopoiesis:the impact of high-throughput sequencing technologies. *Experimental hematology* **39**: 961–8.

Assembly and interrogation of tumor-specific regulatory models reveals master regulators of tumor maintenance and chemosensitivity

Andrea Califano

Center for Computation Biology and Bioinformatics, Columbia University, USA

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

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

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

Reconstructing the Regulatory Network of TB: Transcription Factor Binding Distribution and Properties

Authors: Anna Lyubetskaya, Matthew Peterson, James Galagan.

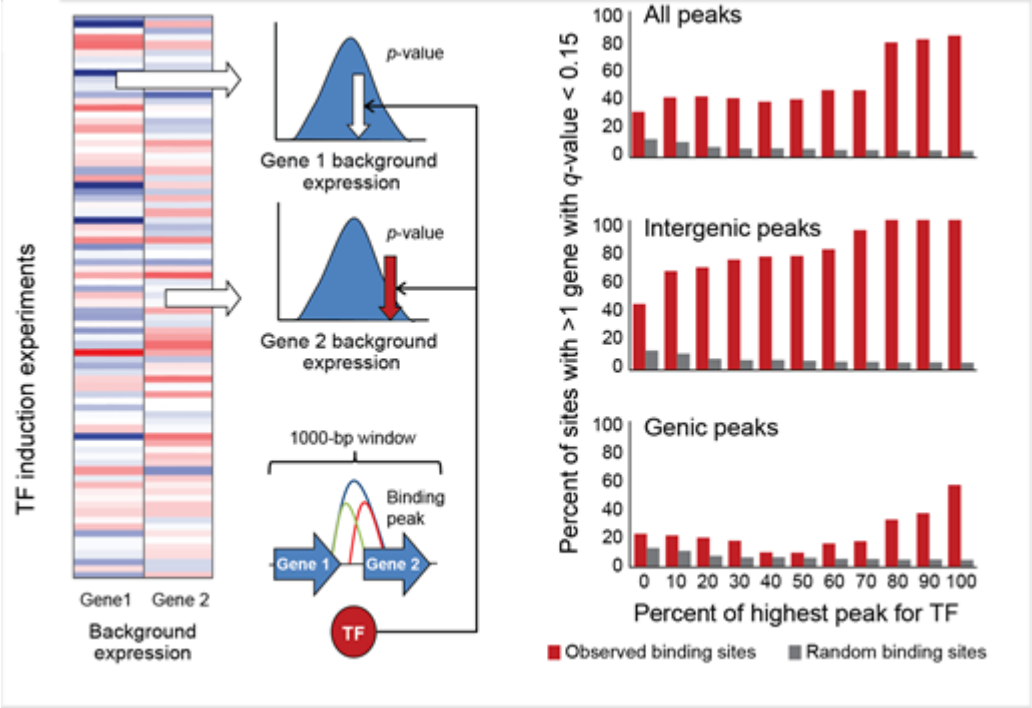
Bioinformatics Program, Boston University, Boston, USA.

E-mail: avl@bu.edu

As a part of the project to identify, validate, and perturb key genes and network interactions predicted to underlie the metabolic adaptations of *M. tuberculosis* and reprogramming of host cells during TB infection, we built a transcriptional network of MTB. So far, we successfully carried out ChIP-Seq experiments for 90 TFs and developed a pipeline to computationally analyze the data. We verified the quality of our data by comparing multiple biological replicates for at least 15 TFs. For 10 key TFs, we also confirmed that our experiments, although carried out in conditions containing oxygen, apply to hypoxic coditions characteristic of MTB infection. While we detected all previously known binding sites for a few well-studied regulators in MTB (KstR and DosR), we also found many more binding instances and significantly extended transcription factor regulons.

In order to validate transcriptional function of predicted binding sites, we carried out complimentary overexpression experiments for all 200 TFs of MTB (replicate experiments were performed for a number of TFs). This data was used to assign a probability of observing the expression level for each gene identified to be bound by a given transcription factor in the overexpression microarrays (Figure 1). For each site, we examined all genes around the site to determine if the overexpression of the corresponding TF significantly altered expression of these genes. Binding sites were validated if any gene in the window displayed an expression level greater than a threshold value after correction for multiple testing.

Applying this method to all sites from analyzed TFs, we could assign a potential regulatory role to 25% of binding sites. Stronger binding sites were more often associated with regulation than weaker sites, suggesting a possible correlation between binding strength and regulatory impact. However, it appeared that clusters of weak binding sites had a stronger regulatory role than weak singletons suggesting cooperative mechanism of interaction. Also, strong binding sites were often located in the proximity of weak sites which suggested the role of weak sites in modulating affinity.



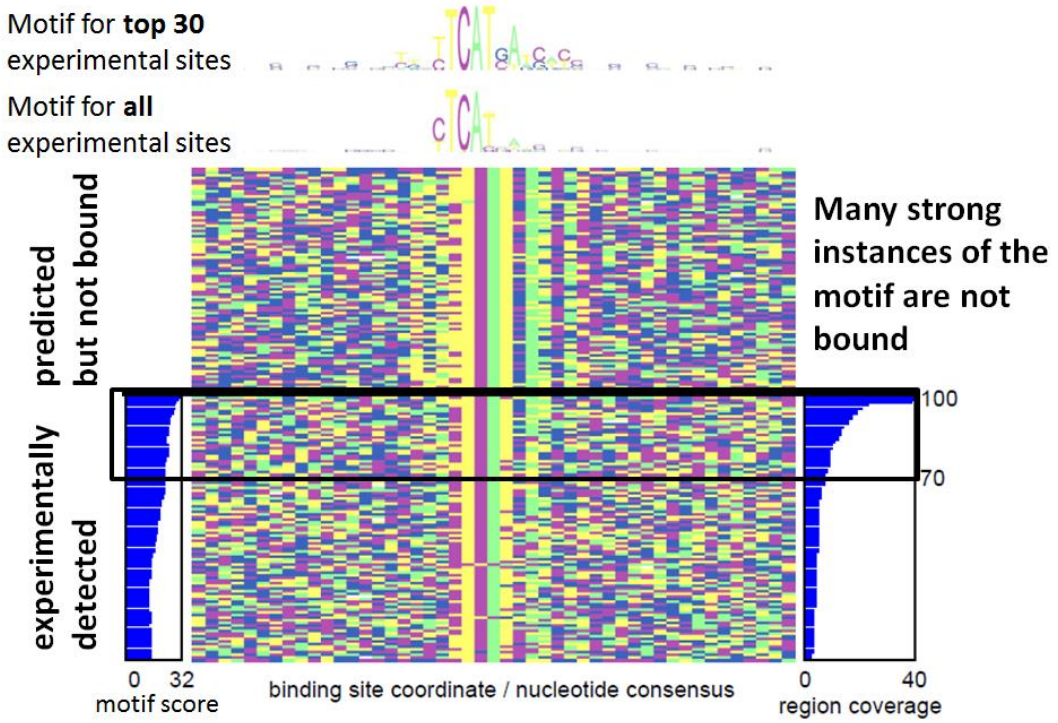
The canonical model of the transcriptional regulation in prokaryotes restricted binding site location to proximal promoter region and suggested that the binding sequence is the main determinant of the binding. The distance between binding sites and associated target genes displayed a pattern partially consistent with expectation: about half of binding sites were located within 1000bp of the start codon of the gene they were predicted to regulate. Most binding sites located in upstream intergenic regions were validated by expression data. However, 76% of binding sites fell into annotated coding regions and a significant proportion could be assigned regulation.

Integration of independent binding and expression datasets allowed us to test which binding site characteristics – binding motif strength, ChIP-Seq coverage, relative location of site and potential target genes, and presence of other binding sites – are essential for assigning regulatory role.

Although a conservative binding motif was found for most transcription factors, only a fraction of motif instances appeared bound in the experiment. The experimentally determined motif for weak binding sites was often a degraded version of the motif detected for the strong binding sites. Some low-affinity binding sites appeared occupied by the transcription factor while many high-affinity binding sites were not.

For example, we detected 100 binding sites for Rv0602c (Figure 2). Site coverage ranges from 40 times above the median to 1 as indicated in the right side of the figure. The strong experimentally detected binding sites are characterized by a TCATGA motif. With coverage, this motif degrades to the TCAT core with some conservation at accesory positions as reflected by the motif score in the left side of the figure. However, if we use

the strong motif to scan the genome, we find many additional instances unbound in the experiment. Interestingly, we find exactly exactly same 13 nulceotides bound in one area of the genome and not bound in another.



By comparing experimental and computational binding site distributions, we defined 'hot' areas of the genome (that were depleted of binding in the experiment despite the existence of motif instances) and 'cold' areas (that were bound by more TFs than expected from the regression model). A nucleoid-associated protein LSR2 with a known role in organizing DNA was associated with these regions, as well as other TFs with no known structural function (for example, Rv0081). Our data suggested that some transcription factors had both distinct regulatory role and significant impact on DNA organization.

The genetics of individuals: why would a mutation kill me, but not you?

Ben Lehner

EMBL-CRG Systems Biology and ICREA, Centre for Genomic Regulation, UPF, Barcelona, Spain

To what extent is it possible to predict the phenotypic differences among individuals from their completely sequenced genomes? We use model organisms (yeast, worms and tumours) to understand when you can, and why you cannot, predict the characteristics of individuals from their genome sequences.

GWAS next generation: identifying mechanisms of action in association studies.

María Rodríguez Martínez, Paola Nicoletti, Gonzalo López, Mukesh Bansal, Yishai Shimoni, Andrea Califano.

Columbia University, 10033 NY, USA.

Genome wide association studies (GWAS) have emerged as a powerful tool for the identification of genetic variants that are associatedwith complex phenotypes and disease.Despite the many newly discovered associations, thevariants identified by these studies typically explain only a small fraction of the heritable component of disease risk[1]. Furthermore,few genetic variants are found within coding regions of genes, and theelucidation of the molecular mechanism by which these loci influencethe phenotype remains challenging. Most loci mapto inter-genic regions of unknown function and, whilesome of them can be connected to nearby genes by linkage disequilibrium, a sizable fraction lie in genomic regions with no clear connection to known disease biology.

The genetic component of complex phenotypes can also arise from a large number of small effect loci[2]. In this case, the heritability would not be due to a single common or rare variant, but rather to combinations of common variants, each one contributing a small additive effect. These combinations can be epistatic interactions among common alleles, or multiple genetic variations that interact through different layers of genomic regulation. Complex phenotypes therefore would have a much more complex genetic architecture due to the joint action of very many loci of small effect [3]. Identifying interactions between multiple loci requires the application of statistical and computational methods that detect patterns of epistasis across the genome. This involves performing genome-wide searches of high order combinations of SNPs or SNPs and genes, and requires testing a large number of hypothesis with often limited sample sizes, leading to a reduced statistical power. Furthermore the computational search becomes unmanageable for more than a few hundredSNPs.

In this work we explore an innovative approach to identify the molecular mechanisms of genetic variants previously associated to disease. We have implemented gVITaMIN (Genetic Variability IdenTifies Missing INteractions), an algorithm that searches forfunctional genetic associations following a two-step approach. First, gVITaMIN searches for direct associations between a locus and gene expression levels. However a SNP can be functionally important for a phenotype without displaying any association with gene expression, therefore in a second step, gVITaMIN searches

for associations between a locus and changes in gene activity. Concretely, we analyze whether aputative locusinfluences the regulatory activity of a transcription factor (TF) over a large set of its target genes (TG). This influence is measured as a difference in the correlation between the TF and its TGs conditioned upon the presence of the variant.

In recent years, a plethora of epigenetic modificationsin the human genome have been characterized and shown to play diverse roles in gene regulation, cellular differentiation and the onset of disease. In particular, regulatory elements such as transcriptional enhancers and silencers, or chromatin marks such as promoters and enhancers, have been shown to play a crucial role in the establishment and maintenance of specific gene regulatory programs. These elements can be perturbed by genetic variants. For instance, mutations in regulatory elements can disrupt or enhance the binding of transcription factors and alter gene expression; polymorphisms that overlap with chromatin marks can prevent regulation through methylation or acetylation and hinder transcription factor binding, etc. In order to integrate this level of genomic regulation, we use to the ENCyclopedia Of DNA Elements (ENCODE) [4, 5] to identify loci that map to characterized functional elements of the human genome. These loci are scored according to their proximity to the genomic element, and linked to the gene or genetic program associated to the functional mark.

Finally, cumulative associations in a particular pathway are likely to pinpoint specific regulatory programs associated with a disease.We therefore search for functional variants associated to a phenotype that cluster on biological units, such as genes or pathways.We combinethe TFs and TGs predicted to be associated to the loci with the genes and genetic programs identified through genomic mapping, and search for enriched pathways on these genes usingGene Set Enrichment Analysis (GSEA) [6]. These pathways are likely to be phenotypically relevant to the physiopathology of the disease and provide insights into the molecular mechanisms underlying them.

We have applied gVITaMIN to the study of genetic variants associated to breast cancer susceptibility. We will present preliminary data identifying an intriguing association between BARD1, a genethat forms a heterodimer with BRCA1, and it is essential for the stability of BRCA1. Several variants at this locus have been reported to be associated with high risk neuroblastoma and colon cancer, suggesting a role in disease that is active across different cancers phenotypes.

BIBLIOGRAPHY

1. Manolio, T.A., *Genomewide association studies and assessment of the risk of disease*. N Engl J Med, 2010. **363**(2): p. 166-76.
2. Valdar, W., et al., *Genome-wide genetic association of complex traits in heterogeneous stock mice*. Nat Genet, 2006. **38**(8): p. 879-87.
3. Eichler, E.E., et al., *Missing heritability and strategies for finding the underlying causes of complex disease*. Nat Rev Genet, 2010. **11**(6): p. 446-50.
4. Ernst, J., et al., *Mapping and analysis of chromatin state dynamics in nine human cell types*. Nature, 2011. **473**(7345): p. 43-9.
5. Ernst, J. and M. Kellis, *Discovery and characterization of chromatin states for systematic annotation of the human genome*. Nat Biotechnol, 2010. **28**(8): p. 817-25.
6. Subramanian, A., et al., *Gene set enrichment analysis: a knowledge-based approach for interpreting genome-wide expression profiles*. Proc Natl Acad Sci U S A, 2005. **102**(43): p. 15545-50.

Alternative and overlapping determinants of transcription start site selection in vertebrate promoters

Boris Lenhard

MRC Clinical Sciences Centre, Imperial College London, United Kingdom

Vertebrate promoters differ with respect to the precision of transcription star site (TSS) selection and the sequence motifs that determine it. The highly precise transcription start sites are found at fixed distances from e.g. TATA box motifs, while most TATA-less promoters allow transcription to start within a broader region. Here we report that TSS selection rules change systematically on subsets of promoters in development and differentiation time courses. The first and most intriguing is the promoter grammar change during maternal to zygotic transition during early embryonic development. We have analysed transcription initiation sites at 1bp resolution in combination with histone modification at core promoter regions at high spatial precision in the course early development of zebrafish. We show that the switch from maternal to zygotic transcriptome is accompanied by a switch between two fundamentally different mechanisms for defining transcription initiation. Upon zygotic transcription activation, the maternal specific W-box motif dependent TSS definition is replaced with a SS|WW dinucleotide enrichment boundary-associated grammar. The two grammars coexist in core promoters of ubiquitously expressed genes in close proximity or in an overlapping fashion and thus enable the continuous expression of these genes in the two very different intracellular environments. The switch in promoter interpretation constitutes a central part of the mechanism for setting up the promoters for the regulation of early development. We show that related, albeit less dramatic systematic changes in TSS selection occur during male spermatogenesis and skeletal muscle differentiation. To ensure gene expression in all stages of these processes, the corresponding promoters must accommodate all the required grammars, often in an overlapping fashion.

Drosophila Pol II core promoters cluster into four classes characterized by distinct sets of motifs, regulatory properties, and nucleosome patterning

Holger Hartmann*, Mark E. L. Heron*, Anja Kiesel*, Lukas Utz, Claudia Gugenmus, and Johannes Söding
(* equal contributions)

Gene Center Munich and Department of Biochemistry, Ludwig Maximillian University, Germany

Core promoters (CPs) are the sites in the genome that recruit the basal transcription machinerie in order to initiate transcription. High-throughput measurements of transcription start site distributions have established the existence of two classes of eukaryotic CPs: Narrow-peaked or “focussed” promoters are usually highly regulated, while broad-peaked or “dispersed” promoters mostly belong to constitutively expressed housekeeping genes. These two classes differ in their motif composition, and it is becoming clear that their motifs influence which combinations of basal transcription factors assemble into a functional preinitiation complex.

We have systematically studied, at the example of Drosophila melanogaster, the link between core promoter elements and the resulting regulatory properties. Our motif discovery method XXmotif finds 12 known and 7 novel, conserved core promoter elements (CPEs). These motifs fall into four groups that tend to co-occur and that characterize four overlapping classes of CPs: (1) strongly regulated, stallable, INR-enriched CP, mostly from developmental genes, (2) highly inducible, TATA-containing CPs, (3) constitutive CPs from housekeeping genes, and (4) very strongly constitutively active CPs, mostly from ribosomal genes. Furthermore, each class has a characteristic dinucleotides profile that is correlated with its nucleosome patterning and 5’-nucleosome-free region. The four CP classes hint at four major, alternative pathways of transcription initiation, each of which uses a different set of basal transcription factors and thereby determines regulatory response. Despite employing different motifs, the same four classes of CPs are likely to exist in humans and other species.

Interplay between chromatin state, regulator binding, and regulatory motifs

Jason Ernst*

UCLA, USA

The regions bound by sequence-specific transcription factors can be highly variable across different cell types, despite the static nature of the underlying genome sequence. This has been partly attributed to changes in chromatin accessibility, but a systematic picture has been hindered by the lack of large-scale datasets. In this talk I will describe our efforts analyzing 456 binding experiments for 119 regulators and 84 chromatin maps generated by ENCODE in six human cell types and relating those to a global map of regulatory motif instances for these factors. We find specific and robust chromatin state preferences for each regulator beyond the previously-reported open-chromatin association, suggesting a much richer chromatin landscape beyond simple accessibility. The preferentially-bound chromatin states of regulators were enriched for sequence motifs of regulators relative to all states, suggesting that these preferences are at least partly encoded by the genomic sequence. Relative to all regions bound by a regulator however, regulatory motifs were surprisingly depleted in the regulator's preferentially-bound states, suggesting additional non-sequence-specific binding beyond the level predicted by the regulatory motifs. Such permissive binding was largely restricted to open-chromatin regions showing histone modification marks characteristic of active enhancer and promoter regions, whereas open-chromatin regions lacking such marks did not show permissive binding. Lastly, the vast majority of co-binding of regulator pairs is predicted by the chromatin state preferences of individual regulators. Overall, our results suggest a joint role of sequence motifs and specific chromatin states beyond mere accessibility in mediating regulator binding dynamics across different cell types.

*Joint work with Manolis Kellis

Predicting regulatory domain boundaries from chromatin immunoprecipitation data

Pawel Bednarz, Bartek Wilczyński

Institute of Informatics, University of Warsaw, Poland

During development of a multicellular organism, cells undergo an orchestrated series of transformations. Through coordinated proliferation and differentiation a complex system of thousands of complementary cells is formed. One of the key aspects of this process is regulation of transcription, allowing different cells to express different sets of proteins leading to variablility in cell morphology and function. In this process, expression of thousands of genes needs to be tightly controlled as misexpression of an important gene in a wrong tissue or at a wrong developmental stage would in many cases lead to developmental defects. This level of control is achieved through the action of transcription factors binding to enhancers, or more generally regulatory elements, leading to very selective activation or repression of their target genes.

Enhancers usually act on target genes' promoters by physically interacting (through co-factor proteins) with the core transcriptional machinery. While in majority of cases of studied enhancers we have a notion of a target gene, it may be diffcult to assign regulatory elements to target genes, especially in the light of recent findings showing enhancer sharing between genes for both humans (Sanyal et al., 2012) and mice (Li et al., 2012).

With the progress of mapping transcription factor binding sites through chromatin immunoprecipitation-based experiments, we are getting closer to having complete maps of regulatory elements in multiple species (Negre et al., 2011). With complementary techniques such as DNAse I hypersensitivity (Thomas et al., 2011) and FAIRE (Giresi et al., 2007) we can get an even more comprehensive picture of the universe of regulatory regions genome- wide, therefore the need for making comprehensive models of regulatory interactions is becoming one of the main challenges in the field.

In a recent work (Wilczyński et al., 2012), we have shown for mesoderm development in Drosphila that given a comprehensive set of Chip-Chip experiments for relevant transcription factors (Zinzen et al., 2009), it is possible to make a computational model making accurate predictions of tissue- and stage-specific gene expression patterns. One of the key elements of the model, indispensible without loss of prediction accuracy, was at least a rudimentary notion of regulatory domains, in this case based on binding of insulator proteins (Negre et al., 2010).

In order to explore the problem of predicting regulatory domain bound- aries more deeply, we have used supervised machine learning approach to make a model of boundary elements using modENCODE data. The model was trained on large scale mapping of chromatin domains (Sexton et al., 2012) and subsequently tested on independent datasets, both high-throughput (Filion et al., 2010) and targeted tests of insulator regions through luciferase assays (Srinivasan and Mishra, 2012). Our model achieves over .80 AUC in cross-validation experiments, generalizes well across different datasets and outperforms other approaches such as Hidden Markov Models (Ernst and Kellis, 2012) or sequence based predictions (Srinivasan and Mishra, 2012).

Prediction of Chromatin State Variability from Genomic Sequence

Luca Pinello[1,2], Jian Xu[2,3,4], Stuart H. Orkin[1,2,3,4], **Guo-Cheng Yuan**[1,2]

[1]Dana-Farber Cancer Institute, [2]Harvard University, [3]Boston Children's Hospital, [4]Howard Hughes Medical Institute

Background

In eukaryotic cells the genome is organized into chromatin. The accessibility of the chromatin varies from one cell-type to another. The resulting constraint on protein-DNA binding provides an important layer of gene regulation. Recent epigenomic studies have uncovered diverse classes of regulatory elements, many of which are located in the regions previously viewed as "junk" DNA, providing strong evidence that chromatin states play a critical role in mediating cell-type specific transcriptional activities. However, the mechanisms underlying the variation of chromatin states remain poorly understood.

We have investigated the role of DNA sequence in mediating the cross cell-type variability of chromatin states with the focus on the histone mark H3K27me3, which mediates cell-type specific gene silencing [1] and plays an important role in the maintenance of cell identity and lineage differentiation [2] [3] . While it is well known that H3K27me3 occupancy is highly enriched at GC rich DNA elements [4], here we focus on distal regions where its recruiting mechanism is less understood [5] .

Results

Genome-wide Characterization of H3K27me3 Plasticity

We obtained a ChIPseq dataset containing H3K27me3 in 19 human cell lines from the ENCODE consortium [6]. The raw-sequence reads data were normalized and mapped to non-overlapping bins of 200bp. The fluctuation of sequence reads can be approximately modeled by a Poisson distribution, which has the distinct property that the mean level is equal to the variance. This motivated us to use the index of dispersion (IOD) statistic to quantify H3K27me3 variability. The Poisson distribution correspond to an IOD value of 1. We selected the top 1% of bins with highest IOD values and referred to those as the most variable regions (MVR) (Figure 1, green dots), whereas the bottom 1% of bins were referred to as the least variable regions (LVR) (Figure 1, red dots).

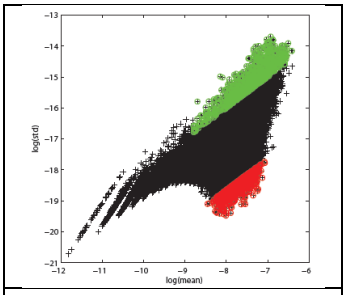


Figure 1. Definition of the MVR.

To test whether the variation of H3K27me3 was indeed associated with cell-type specific gene regulation, we investigated the correlation between the dynamic change of H3K27me3 occupancy and the expression levels of the neighboring genes. We found that, for most regions, there is a significant correlation between H3K27me3 density and gene expression level.

Prediction of H3K27me3 Plasticity from Genomic Sequences

The genome-wide distribution of H3K27me3 is regulated by both sequence dependent and independent mechanisms. On one hand, previous studies have identified a number of DNA sequence features associated with H3K27me3, including CpG islands [4], transcription factor sequence motifs [7] [8], short RNA hairpins [9], and lincRNA [10] [11] . On the other hand, existing H3K27me3 can be recognized by chromatin regulators thereby propagating in a self-enhancing manner. Previous studies have been focused on a specific cell-type, whereas to what extent the DNA sequence regulates the overall variability remains poorly understood. Of note, while the prediction of cell-type specific changes requires additional factors than sequence information, which is cell-type independent, it remains possible to predict the overall variability with accuracy as shown below.

We applied N-score, a computational method previously developed for nucleosome positioning prediction [12], to predict the location of MVRs from the underlying genomic sequences, using LVRs as negative control. We evaluated the model performance by cross-validation and obtained an accuracy of AUC = 0.82 (Figure 2). We then applied our model over running windows across the entire genome, and compared the predicted variability with ChIPseq data. The genome-wide correlation with experimental data is $\rho = 0.28$.

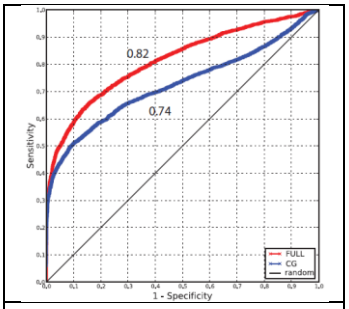


Figure 2. ROC for prediction of MVR locations from DNA sequence.

Distal MVRs Are Regulated by Cell-type Specific Transcription Factors

Next we focused on the MVRs in distal regions, which have recently been found to contain many important enhancer elements. We compared with a publicly available dataset of genome-wide enhancers in 9 ENCODE cell lines [13], and found that the distal MVRs are highly enriched with enhancers present in at least one cell line (p-value < 2.2E-16).

Compared to proximal MVRs, the distal MVRs tend to have lower mean and variance. Interestingly, the H3K27me3 density at distal MVRs appear to be bimodal: while the value is comparable to background level in most cell lines, it is significantly higher in one or two specific cell-types, suggesting an important role of cell-type specific regulators in their recruitment.

Since the distal MVRs are markedly cell-type specific, we searched for candidate TFs that may a role in Polycomb group (PcG) recruitment in cell-type specific manner. For each cell-type, we ranked the MVRs according to the z-score and selected the top ranking ones as the cell-type specific subset. We searched for known transcription factor (TF) motifs that are over-represented in each cell-type specific MVRs while using the rest as the background. For most cell-types, we were able to identify a small number motifs that are highly significantly over-represented.

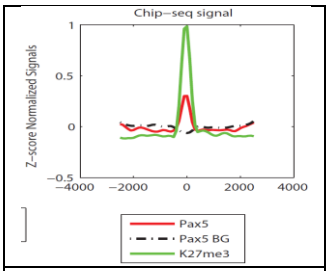


Figure 3. PAX5 is enriched in lymphoblastsoid-specific MVRs.

As an example, we found that the PAX5 motif is highly enriched in the lymphoblastoid cell lines (GM12878 and GM06990). Furthermore, the expression level of PAX5 is also higher in this cell-type than others, consist with the known role of PAX5 in B-cell development. Indeed, a role of Polycomb recruitment of PAX5 has previously been identified. To test whether PAX5 may facilitate PcG binding in this cell line, we tested its colocalization with H3K27me3 by using public ChIPseq data. We found that PAX5 and H3K27me3 indeed colocalize at these MVRs (Figure 3). Consistent with a gene silencing role, the target gene expressions are lower in these cell lines that the rest.

A Role of the TAL1 in Regulating H3K27me3 Recruitment in Erythroid Precursors

Next, we investigated whether the computational strategy discussed above may be useful for prediction of novel PcG recruiting factors in less well-characterized systems. In a recent study, we have characterized the genome-wide chromatin states in erythroid precursors (ProE) using primary human cell lines, and found that enhancer mediated gene activities are responsible for developmental-stage selection [14]. Using the same strategy as described above, we integrated our H3K27me3 ChIPseq data for ProE together with those obtained from ENCODE, and identified a subset of distal MVRs that are specific to ProE.

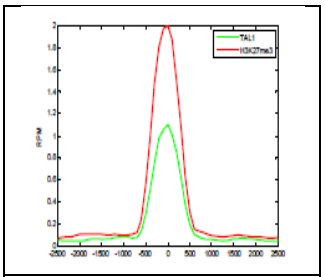


Figure 4. TAL1 is enriched in ProE specific MVRs.

In order to identify ProE-specific PcG recruiting factors, we searched for enriched TF motifs in the ProE-specific distal MVRs. One of the most enriched motifs corresponds to TAL1 (p-value = 5.6E-37). This is surprising because TAL1 is a well-characterized activator that is required for erythroid development. Although a possible role in repression has recently been suggested [15], a mechanistic understanding is still lacking. Our analysis suggests that TAL1 may play a role in PcG recruitment thereby repressing the target genes. We then examined the TAL1 ChIPseq data around the distal MVRs, and indeed found significant TAL1 binding signal (Figure 4). Furthermore, gene expression data analysis showed that the expression level of the target genes are expressed at a

lower level compared to the overall TAL1 target genes. These results support a role of TAL1 in orchestrating PcG recruitment during erythroid development.

Conclusions

We have developed a systematic approach to investigate the mechanisms regulating chromatin state variability and applied it to H3K27me3. We found that the MVRs can be well-predicted by the underlying DNA sequences. Furthermore, the distal MVRs cannot be explained by GC content but are enriched for cell-type specific TF motifs. Using this approach, we found that the erythroid master regulator TAL1, which is commonly known as an activator, can also play a role in gene repression by targeted recruitment of Polycomb complexes. Our approach is generally applicable to other epigenetic marks.

1. Francis NJ, Kingston RE: **Mechanisms of transcriptional memory.** *Nat Rev Mol Cell Biol* 2001, **2:**409-421.
2. Boyer LA, Plath K, Zeitlinger J, Brambrink T, Medeiros LA, Lee TI, Levine SS, Wernig M, Tajonar A, Ray MK, et al: **Polycomb complexes repress developmental regulators in murine embryonic stem cells.** *Nature* 2006, **441:**349-353.
3. Bernstein BE, Mikkelsen TS, Xie X, Kamal M, Huebert DJ, Cuff J, Fry B, Meissner A, Wernig M, Plath K, et al: **A bivalent chromatin structure marks key developmental genes in embryonic stem cells.** *Cell* 2006, **125:**315-326.
4. Mendenhall EM, Koche RP, Truong T, Zhou VW, Issac B, Chi AS, Ku M, Bernstein BE: **GC-rich sequence elements recruit PRC2 in mammalian ES cells.** *PLoS Genet* 2010, **6:**e1001244.
5. Arnold P, Scholer A, Pachkov M, Balwierz PJ, Jorgensen H, Stadler MB, van Nimwegen E, Schubeler D: **Modeling of epigenome dynamics identifies transcription factors that mediate Polycomb targeting.** *Genome Res* 2013, **23:**60-73.
6. Dunham I, Kundaje A, Aldred SF, Collins PJ, Davis CA, Doyle F, Epstein CB, Frietze S, Harrow J, Kaul R, et al: **An integrated encyclopedia of DNA elements in the human genome.** *Nature* 2012, **489:**57-74.
7. Liu Y, Shao Z, Yuan GC: **Prediction of Polycomb target genes in mouse embryonic stem cells.** *Genomics* 2010, **96:**17-26.
8. Ku M, Koche RP, Rheinbay E, Mendenhall EM, Endoh M, Mikkelsen TS, Presser A, Nusbaum C, Xie X, Chi AS, et al: **Genomewide analysis of PRC1 and PRC2 occupancy identifies two classes of bivalent domains.** *PLoS Genet* 2008, **4:**e1000242.
9. Kanhere A, Viiri K, Araujo CC, Rasaiyaah J, Bouwman RD, Whyte WA, Pereira CF, Brookes E, Walker K, Bell GW, et al: **Short RNAs are transcribed from repressed polycomb target genes and interact with polycomb repressive complex-2.** *Mol Cell* 2010, **38:**675-688.
10. Khalil AM, Guttman M, Huarte M, Garber M, Raj A, Rivea Morales D, Thomas K, Presser A, Bernstein BE, van Oudenaarden A, et al: **Many human large intergenic noncoding RNAs associate with chromatin-modifying complexes and affect gene expression.** *Proc Natl Acad Sci U S A* 2009, **106:**11667-11672.
11. Margueron R, Reinberg D: **Chromatin structure and the inheritance of epigenetic information.** *Nat Rev Genet* 2010, **11:**285-296.
12. Yuan GC, Liu JS: **Genomic sequence is highly predictive of local nucleosome depletion.** *PLoS Comput Biol* 2008, **4:**e13.
13. Ernst J, Kheradpour P, Mikkelsen TS, Shoresh N, Ward LD, Epstein CB, Zhang X, Wang L, Issner R, Coyne M, et al: **Mapping and analysis of chromatin state dynamics in nine human cell types.** *Nature* 2011, **473:**43-49.
14. Xu J, Shao Z, Glass K, Bauer DE, Pinello L, Van Handel B, Hou S, Stamatoyannopoulos JA, Mikkola HK, Yuan GC, Orkin SH: **Combinatorial assembly of developmental stage-specific enhancers controls gene expression programs during human erythropoiesis.** *Dev Cell* 2012, **23:**796-811.
15. Van Handel B, Montel-Hagen A, Sasidharan R, Nakano H, Ferrari R, Boogerd CJ, Schredelseker J, Wang Y, Hunter S, Org T, et al: **Scl represses cardiomyogenesis in prospective hemogenic endothelium and endocardium.** *Cell* 2012, **150:**590-605.

Rotational positioning of regulatory elements within nucleosomes

Jan Hapala[1], jan@hapala.cz,
Edward N. Trifonov[1,2], trifonov@research.haifa.ac.il,

[1]Faculty of Science and CEITEC – Central European Institute of Technology,
Masaryk University, Kamenice 5, 62500 Brno, Czech Republic
[2]Genome Diversity Center, Institute of Evolution,
University of Haifa, Mount Carmel, 31905 Haifa, Israel

1 Regulation of gene expression depends on the rotational positioning

One of the key factors in the regulation of gene expression is binding of transcription factors to their response elements. This binding is well known to depend on rotational orientation of the binding sites in the nucleosome [6, 7, 12, 9, 1].

Typically, the binding occurs if the binding site is exposed. The transcription factor binds to the DNA with the highest affinity when the recognition sites are oriented away from the histone octamer. For example p53 binds preferentially to the nucleosomes when the minor grooves of the recognition sequences are oriented outwards [9]. Another example is glucocorticoid receptor which favors the binding sites positioned in the major grooves facing out [6]. In both cases the transcription factor binding decreases when rotational positioning of the recognition sites is changed, being abolished when the sites face the histone octamer [6, 9].

2 Computational mapping of the nucleosomes

The rotational setting can be determined by any of few available techniques, which allow a single-base resolution mapping of the nucleosomes on DNA

[3, 11, 2]. We used the DNA bendability matrix derived by [3] to study the rotational positioning of TATA boxes and splice junctions.

This technique had been tested on the set of nucleosome DNA sequences experimentally mapped with high accuracy, including the crystallized nucleosome data. The test demonstrated ± 1 base fit [3] to the experimental positions.

3 Rotational positioning of the TATA boxes

For the analysis of the TATA boxes we extracted[5] DNA sequences from the Eukaryotic Promoter Database [8] and mapped nucleosomes on these sequenced synchronized around the TATA box.

Our results show that the nucleosome DNA sequence harboring the TATA box encodes alternative rotational positions for the same piece of DNA. This may serve for switching the gene activity on and off.

4 Rotational positioning of the splice junctions

When we applied[4] this approach to DNA sequences containing splice junctions from five different species[10], we found the junctions to be preferentially located within nucleosomes. Moreover, the orientation of guanine residues at the GT- and AG-ends of introns within the nucleosomes are such that the guanines are positioned nearest to the surface of histone octamers, 3 and 4 bases upstream from the local DNA pseudo-dyads passing through minor grooves oriented outwards. Since the guanine residues are the most vulnerable to spontaneous damage within the cell (primarily, depurination and oxidation) such positioning of the splice junctions minimizes the damage that is caused by free radicals and highly reactive metabolites.

References

[1] P Blomquist, S Belikov, and O Wrange. Increased nuclear factor 1 binding to its nucleosomal site mediated by sequence-dependent DNA structure. *Nucleic Acids Research*, 27(2):517–525, 1999.

[2] Kristin Brogaard, Liqun Xi, Ji-Ping Wang, and Jonathan Widom. A map of nucleosome positions in yeast at base-pair resolution. *Nature*, 486(7404):496–501, 2012.

[3] Idan Gabdank, Danny Barash, and Edward N. Trifonov. Single-base resolution nucleosome mapping on DNA sequences. *Journal of Biomolecular Structure & Dynamics*, 28(1):107–121, 2010.

[4] Jan Hapala and Edward N. Trifonov. High resolution positioning of intron ends on the nucleosomes. *Gene*, 489(1):6–10, 2011.

[5] Jan Hapala and Edward N. Trifonov. Nucleosomal TATA-switch: competing orientations of TATA on the nucleosome. *Gene*, submitted in 2013.

[6] Q Li and O Wrange. Accessibility of a glucocorticoid response element in a nucleosome depends on its rotational positioning. *Molecular and Cellular Biology*, 15(8):4375–4384, 1995.

[7] Q Li, O Wrange, and P Eriksson. The role of chromatin in transcriptional regulation. *International Journal of Biochemistry & Cell Biology*, 29(5):731–742, 1997.

[8] RC Perier, T Junier, and P Bucher. The Eukaryotic promoter database EPD. *Nucleic Acids Research*, 26(1):353–357, 1998.

[9] Geetaram Sahu, Difei Wang, Claudia B. Chen, Victor B. Zhurkin, Rodney E. Harrington, Ettore Appella, Gordon L. Hager, and Akhilesh K. Nagaich. p53 binding to nucleosomal DNA depends on the rotational positioning of DNA response element. *Journal of Biological Chemistry*, 285(2):1321–1332, 2010.

[10] S Saxonov, I Daizadeh, A Fedorov, and W Gilbert. EID: the Exon-Intron Database - an exhaustive database of protein-coding intron-containing genes. *Nucleic Acid Research*, 28(1):185–190, 2000.

[11] Michael Y. Tolstorukov, Vidhu Choudhary, Wilma K. Olson, Victor B. Zhurkin, and Peter J. Park. nuScore: a web-interface for nucleosome positioning predictions. *Bioinformatics*, 24(12):1456–1458, 2008.

[12] JM Wong, Q Li, BZ Levi, YB Shi, and AP Wolffe. Structural and functional features of a specific nucleosome containing a recognition element for the thyroid hormone receptor. *EMBO Journal*, 16(23):7130–7145, 1997.

Does intragenic DNA methylation determine differential exon expression?

Meromit Singer[1]*, **Idit Kosti**[2]*, Lior Pachter[1,3,4] and Yael Mandel-Gutfreund[2]

[1] Computer Science Division, University of California, Berkeley, CA, USA, [2] Faculty of Biology, Technion - Israel Institute of Technology, Technion City, Haifa, Israel , [3] Department of Mathematics, University of California, Berkeley, CA, USA , [4]Department of Molecular & Cell Biology, University of California, Berkeley, CA, USA

* Equal contribution

**To whom correspondence should be addressed: yaelmg@tx.technion.ac.il

Background

DNA methylation is an important epigenetic marker associated with the regulation of gene expression in eukaryotes. While promoter methylation is relatively well-characterized as a gene silencer in vertebrates, the role for intragenic methylation remains unclear. The genome-wide location of intragenic DNA methylation was determined in many eukaryotic species [1], along with analyses of messenger RNA. A recent study suggests that DNA methylation affects exon recognition and is influenced by the GC architecture of the exon and flanking introns [2]. In this study we investigate the role of DNA methylation in the exons and their flanking introns based on DNA methylation and expression data.

Results

Our data consists of 32000 exons with RNA-seq expression and BS-seq methylation data from Human Fibroblast cell-line IMR90 [1]. Further we extract four intronic regions of length 200 bp flanking each exon representing the middle of upstream intron, the immediate region upstream intron region, immediate intron region downstream and the middle of the downstream intron.

Strikingly we noticed a significant difference in the methylation pattern of intronic regions flanking highly methylated exons versus low methylated exons. Specifically, the highly methylated exons (figure 1A) were found to be significantly more methylated than their intronic surroundings while the low methylated exons (figure 1B) showed the opposite pattern, where the flanking introns had higher methylation levels.

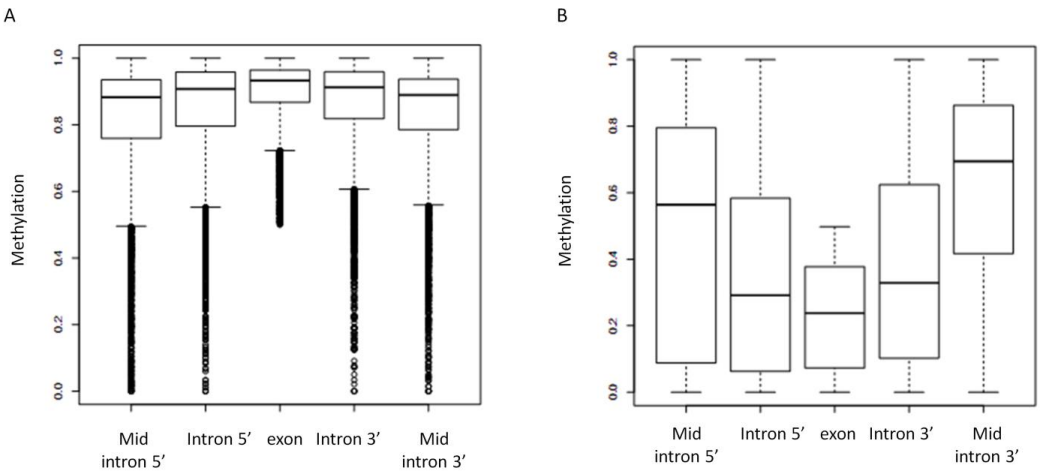


Figure 1: Methylation levels in exons and flanking introns for high [A] and low [B] methylated exons.

Furthermore, the highly methylated exons were highly expressed while low methylated exons were on general weakly expressed. Interestingly, in both top and lowest expressed exons we notice two distinct patterns of methylation (we name Peak and Dip), suggesting two alternative mechanisms relating intragenic DNA methylation to exon expression. Overall, the different methylated patterns were not correlated with either the GC content or the evolutionary conservation of the exons and their flanking introns.

Last, we explore the relation between promoter methylation and exon methylation and expression. While we did not detect a linear correlation between exon/intron methylation levels and the promoter methylation, we show that highly methylated exons tend to have higher promoter methylation and accordingly lower expression.

Conclusions

Consistent with recent studies [2], this study reinforces that the differential methylation level of the exon and its intronic surroundings can dictate exon faith. Specifically we show a positive correlation between intragenic methylation and exon expression. Overall our results strongly suggest that exon expression is influenced by the local methylation state, independent of the overall expression of the gene and the methylation status of its promoter.

References

[1] Lister R. et al, *Human DNA methylomes at base resolution show widespread epigenomic differences*, Nature 462, 315-322.

[2] Gelfman S. et al, *DNA-methylation effect on cotranscriptional splicing is dependent on GC architecture of the exon-intron structure,* Genome research, March 15, 2013

Survival of the quickest – Identification of time-optimal regulatory strategies of metabolism in *Escherichia coli*

Christoph Kaleta (christoph.kaleta@uni-jena.de)

Research Group Theoretical Systems Biology, University of Jena, Leutragraben 1, 07743 Jena, Germany

1 Introduction

While the ability of microorganisms to adapt to specific environmental conditions has been studied with much detail in the past, specific mechanisms to cope with changing environments only recently have received increasing attention (Buescher et al., 2012; Schuetz et al., 2012; Wessely et al., 2011). In two recent studies, work from my group has focused on the identification of specific regulatory strategies that allow microorganisms to reduce their response time when facing a change in nutritional conditions (Wessely et al., 2011; Bartl et al., 2013). Being able to quickly adjust fluxes through metabolic pathways is of central importance in order to reduce lag times in case of deprivation of an essential nutritional constituent or during major growth-transitions such as the exit from stationary phase (Geisel et al., 2011; Geisel, 2011; Schuetz et al., 2012).

2 Results

A minimal regulatory strategy

In the first study (Wessely et al., 2011), we investigated the coexpression of enzymes belonging to the same pathway on the level of the entire metabolism in *Escherichia coli*. While we found a large number of subsystems of metabolism in which most pathways showed a strong coexpression we also identified several subsystems in which pathways appeared not to be co-regulated. In order to understand how metabolic pathways could be controlled without a consistent regulation across all enzymes, we used dynamic optimization to identify a regulatory strategy that allows to precisely control the flux through a metabolic pathway with a minimal amount of transcriptional regulatory interactions. To this end, we studied a prototypical example pathway comprising five reactions governed by irreversible Michaelis-Menten-Kinetics that convert a buffered substrate into a product that is drained with varying dilution rates in the course of the simulation. Using dynamic optimization we searched for a time-course of the enzymes that maintains the concentration of the product in a narrow predefined range while minimizing a weighted sum of initial enzyme concentrations and the amount of regulation. The amount of regulation was measured as the deviation of enzyme concentrations from their initial value.

The results of the optimization showed that, in case of a low weight of initial enzyme concentrations, it is optimal to transcriptionally control a metabolic pathway only in the initial and terminal step of the pathway. We call this program of regulation "sparse transcriptional regulation". With an increasing weight of initial enzyme concentrations, we observed a shift from sparse transcriptional regulation to the regulation of all enzymes within the pathway, which we called

"pervasive transcriptional regulation". To test the predictions of the optimization approach, we analyzed the pathway-position dependent occurrence of transcriptional regulatory interactions. We could confirm that within the subsystems of metabolism in which we didn't find a co-expression of all enzymes within a pathway, there was a significant increase in the frequency of transcriptional regulatory interactions at the beginning and end of pathways. Moreover, we observed a sparse transcriptional regulation in particular for pathways consisting of lowly abundant enzymes.

We explained the optimality of these different programs of pathway activation by a trade-off between response time and protein cost. A sparse transcriptional regulation of a pathway allows the organism to quickly adjust the flux through a pathway since only the concentration of key enzymes needs to be adjusted. However, it entails a high protein cost since enzymes at intermediate positions within these pathways are expressed constitutively. In contrast, a pervasive transcriptional regulation entails a slow response time since the concentrations of all enzymes need to be adjusted but proteins are only produced if they are needed. Thus, depending on the requirement for a rapid response or a minimization of protein cost, either a sparse or pervasive transcriptional control is optimal. In consequence, it is optimal to sparsely regulate metabolic pathways with a low protein cost (e.g. in co-factor synthesis) while it is optimal to pervasively control metabolic pathways with a high protein cost (e.g. in amino acid biosynthesis). A notable exception is the pentose phosphate pathway that has a high protein cost but shows a pattern of sparse transcriptional regulation. This pathway produces reduction equivalents in the form of NADPH that are required by a large number of other pathways. In consequence, being able to quickly adjust the flux through the pentose phosphate pathway appears to outweigh the high protein cost.

Optimal programs of pathway activation

In a second study, we analyzed how a pathway is optimally activated in the light of limitations of the cellular protein synthesis capacity (Bartl et al., 2013). To this end, we studied a simple metabolic pathway consisting of four enzymatic steps that convert a buffered substrate into a product that is limiting for growth. Starting from an initially inactive pathway, we investigated how the individual enzymes are optimally activated given two constraints on enzyme synthesis to minimize the time until a resumption of growth. These two constraints are a maximal synthesis rate of the individual enzymes and the free protein synthesis capacity that puts an upper limit on the total amount of enzymes that can be synthesized at the same time. While the former constraint is strongly influenced by the amount of protein that is required, the free protein synthesis capacity is influenced by the number of free ribosomes. With an increasing enzyme synthesis rate relative to free protein synthesis capacity, we found a shift from the optimality of a simultaneous activation of all enzymes over a sequential activation of groups of enzymes to a sequential activation of individual enzymes within the pathway. Thus, we found, in contrast to previous works, that a sequential activation of enzymes is only optimal in the case of high protein costs. Moreover, we found that large differences in the abundance of proteins within a pathway lead to the optimality of a accelerated activation of highly abundant enzymes while the induction of lowly abundant enzymes are delayed.

In order to validate the predictions by the optimization approach, we studied the operonic structure of a large number of metabolic pathways across 550 prokaryotes from the MicroCyc collection of metabolic pathways (Vallenet et al., 2013). The operonic organization of the genes of a metabolic pathway allowed us to deduce the particular regulatory program that is used for its control since enzymes within an operon are activated almost concomitantly. Thus, in accordance with our predictions we expected operon sizes to decrease with increasing protein abundance and to increase with increasing protein synthesis capacity. Moreover, we expected highly abundant enzymes within a pathway to be more often coexpressed with earlier enzymes of the same path-

way while we expect lowly abundant enzymes to be more often coexpressed with later steps of a metabolic pathway. Of the 99 pathways with sufficient data of all organisms across the MicroCyc collection, we could confirm for 21 that the dependence between protein abundance and protein synthesis capacity followed our predictions. Notably we found only two cases in which we found significant correlations opposite to our predictions. Moreover, we could confirm that highly abundant proteins are more often coexpressed with earlier enzymes of a pathway while lowly abundant enzymes tend to be coexpressed with later steps.

3 Discussion

These two studies show that with an increasing abundance of proteins within a pathway, the complexity of the transcriptional regulatory programs used for its control drastically increases. For pathways with lowly abundant proteins a focused regulation of key steps is optimal, while a high protein cost entails the optimality of distinct activation times of individual enzymes. Apart from the identification of optimal programs of pathway control, results from these studies are also of importance to identify pathways that are differentially expressed between conditions and for our understanding of the evolution of operons. Relating to the identification of differentially expressed genes, our results imply that, depending on the transcriptional regulatory program used to control a metabolic pathway, a differential activity of a metabolic pathway might only be obvious from changes in the first and/or terminal step. Concerning the evolution of operons, the abundance of enzymes within a pathway as well as the the capacity of the protein synthetic machinery can change in the evolutionary history of an organism. Thus, also the optimal operonic organization of the enzymes of a pathway changes which could partially explain the high evolutionary plasticity of operons even between closely related species (Price et al., 2006).

References

M. Bartl, M. Kötzing, S. Schuster, P. Li, and C. Kaleta. Dynamic optimization identifies optimal programs for pathway regulation in prokaryotes. Nat Comm, 2013. revised version submitted.

J. M. Buescher, W. Liebermeister, M. Jules, M. Uhr, J. Muntel, E. Botella, B. Hessling, R. J. Kleijn, L. L. Chat, F. Lecointe, U. Mäder, P. Nicolas, S. Piersma, F. Rügheimer, D. Becher, P. Bessieres, E. Bidnenko, E. L. Denham, E. Dervyn, K. M. Devine, G. Doherty, S. Drulhe, L. Felicori, M. J. Fogg, A. Goelzer, A. Hansen, C. R. Harwood, M. Hecker, S. Hubner, C. Hultschig, H. Jarmer, E. Klipp, A. Leduc, P. Lewis, F. Molina, P. Noirot, S. Peres, N. Pigeonneau, S. Pohl, S. Rasmussen, B. Rinn, M. Schaffer, J. Schnidder, B. Schwikowski, J. M. V. Dijl, P. Veiga, S. Walsh, A. J. Wilkinson, J. Stelling, S. Aymerich, and U. Sauer. Global network reorganization during dynamic adaptations of *Bacillus subtilis* metabolism. Science, 335(6072):1099–1103, Mar 2012.

N. Geisel. Constitutive versus responsive gene expression strategies for growth in changing environments. PLoS One, 6(11):e27033, 2011.

N. Geisel, J. M. G. Vilar, and J. M. Rubi. Optimal resting-growth strategies of microbial populations in fluctuating environments. PLoS One, 6(4):e18622, 2011.

M. N. Price, A. P. Arkin, and E. J. Alm. The life-cycle of operons. PLoS Genet, 2(6):e96, Jun 2006.

R. Schuetz, N. Zamboni, M. Zampieri, M. Heinemann, and U. Sauer. Multidimensional optimality of microbial metabolism. Science, 336(6081):601–604, May 2012.

D. Vallenet, E. Belda, A. Calteau, S. Cruveiller, S. Engelen, A. Lajus, F. L. Fèvre, C. Longin, D. Mornico, D. Roche, Z. Rouy, G. Salvignol, C. Scarpelli, A. A. T. Smith, M. Weiman, and C. Médigue. Microscope–an integrated microbial resource for the curation and comparative analysis of genomic and metabolic data. Nucleic Acids Res, 41(Database issue):D636–D647, Jan 2013.

F. Wessely, M. Bartl, R. Guthke, P. Li, S. Schuster, and C. Kaleta. Optimal regulatory strategies for metabolic pathways in *Escherichia coli* depending on protein costs. Mol Syst Biol, 7:515, 2011.

EPSILON: localized networks for eQTL prioritization

Lieven P.C. Verbeke[1], Piet Demeester[1], Jan Fostier[1], and Kathleen Marchal[2,3]

[1] *IBCN - iMinds, Ghent University, Belgium, lieven.verbeke@intec.ugent.be*
[2] *Department of Microbial and Molecular Systems, KU Leuven, Belgium*
[3] *Department of Plant Biotechnology and Bioinformatics, Ghent University, Belgium*

Abstract

When genomic data is associated with gene expression data, the resulting expression quantitative trait loci (eQTL) will very likely span multiple genes. eQTL prioritization techniques can be used to select the most likely causal gene affecting the expression of a target gene from a list of candidates. As an input, these techniques use physical interaction networks that often contain highly connected genes and unreliable or irrelevant interactions that can interfere with the prioritization process. We present EPSILON, a framework for eQTL prioritization that mitigates the effect of highly connected genes and unreliable interactions. We tested the new method on eQTL data sets derived from yeast data. A physical interaction network was constructed and each eQTL in each data set was prioritized using the EPSILON approach: first a local network was constructed using a k-trials shortest path algorithm, followed by the calculation of a network-based similarity measure. We found that using a local network significantly increased prioritization performance in terms of predicted knockout pairs when compared to using exactly the same network similarity measures on the global network. EPSILON performed on par or better than two alternative eQTL prioritization approaches, ITM-Probe and eQED.

1 Introduction

Due to linkage disequilibrium and the spacing of the genetic markers on the genome, genetic markers represent a region on a chromosome that covers multiple genes rather than a single gene. The variability in expression of the genes found to be associated with an eQTL (here referred to as *target* genes) is most likely caused by a mutation in a single gene located on the eQTL (the *causal* gene). Gene prioritization or refinement methods are needed to distinguish the causal gene from a list of candidate causal genes.

A relatively small number of techniques were developed to tackle the rather specific eQTL prioritization task. All eQTL prioritization methods have in common that they use a physical interaction network to define a similarity measure between a target gene and a set of candidate causal genes. Tu *et al.* (2006) developed a method based on random walks in a physical interaction network, an approach later refined by Suthram *et al.* (2008), who extended the random walk idea with an electric circuit

analogy. Voevodski *et al.* (2009) applied the PageRank algorithm to develop a gene affinity measure, and Stojmirović and Yu (2012) used the mathematical modeling of information flow in a network to rank candidate genes.

Stojmirović and Yu (2012) suggest localizing the network, i.e. excluding distant genes from the network that connects an origin (the target gene) with a set of destinations (the candidate causal genes), prior to analysis in order to better reflect the biological context. Otherwise, results of e.g. gene prioritization will be highly dependent on the node degree of the genes in the network. Simply removing genes from the network with a node degree exceeding an arbitrary threshold, or heuristically downweighting the importance of relations based on the number of connections, risks removing useful genes or important relations (Zotenko *et al.*, 2008). To handle both localization and prioritization simultaneously, we present EPSILON.

2 EPSILON framework

The EPSILON method contains two steps, which are applied to each association found: (1) construct, from an existing global interaction network, a local sub-network that connects the candidate causal genes covered by an eQTL with the target gene and (2) calculate a similarity measure that expresses the functional similarity between the target gene and a candidate causal gene. As input, the results of an eQTL association analysis are used.

To restrict the network around a set of candidate causal genes and a target gene, a shortest/cheapest path approach is applied. All interactions are assigned a cost, and an optimal path from each candidate to the target was found using the Dijkstra algorithm. All genes and interactions that were found on such a shortest path were included in the sub-network. Furthermore, it was investigated if enlarging this neighborhood could improve the prioritization results. This was achieved by k times considering if an alternative shortest path exists, that is different from any previously found path.

Once the local network connecting all candidate causal genes with the target gene is constructed, the EPSILON framework requires the calculation of a network similarity measure between the target gene and all candidates to assess their functional relatedness. In principle, any network-based similarity measure could be integrated. Several authors (e.g. Tu *et al.* (2006), Suthram *et al.* (2008), Shih and Parthasarathy (2012)) propose a random walk (RW) approach, in which a random walk is initiated a very high number of times from a candidate causal gene, and it is measured how many times a random walker is found in the target gene.

Next to integrating random walks in EPSILON, we investigated kernels calculated on graph nodes as an alternative similarity measure. These kernels are an attractive tool for uncovering relations in large networks (Fouss *et al.*, 2006). In this study, we evaluated two well-known kernels, the Regularized Commute-Time (RCT) kernel and the Laplacian Exponential Diffusion (LED) kernel.

3 Results and Discussion

We evaluated EPSILON, a k-trials shortest path network construction method combined with random walk and kernel-based similarity measures, using a gold standard data set derived from a yeast knockout compendium. We applied three commonly used association techniques to the SNP and expression data (Saccharomyces cerevisiae) of Brem and Kruglyak (2005): non-parametric regression, mixed models and elastic net regression. An interaction network was constructed using public databases, containing protein-protein interactions, transcription factors with targets and phosphorylation interactions.

We were able to show that our approach, outperformed random assignment and a shortest path reference method. More interestingly, the global network analogues of the network similarity measures too were outperformed significantly ($p < 10^{-5}$), clearly showing the added value of using local over global networks. We assume that constraining the global network to a local neighborhood around the target gene and all candidate causal genes is effectively reducing the disturbing impact of hubs and promiscuous genes. EPSILON was compared to two other methods, ITM Probe and eQED. We found that EPSILON performed as well or better than ITM Probe. EPSILON clearly outperformed eQED, be it using a reduced network because eQED could not deal with the phosphorylation interactions present in the global network.

4 Acknowledgments

This work is supported by: (1) Ghent University Multidisciplinary Research Partnership 'Bioinformatics: from nucleotides to networks', (2) Fonds Wetenschappelijk Onderzoek-Vlaanderen (FWO) G.0428.13N and (3) Katholieke Universiteit Leuven funding: PF/10/010 (NATAR).

References

Brem, R. B. and Kruglyak, L. (2005). The landscape of genetic complexity across 5,700 gene expression traits in yeast. *Proceedings of the National Academy of Sciences of the United States of America*, **102**(5), 1572–7.

Fouss, F., Yen, L., Pirotte, A., and Saerens, M. (2006). An Experimental Investigation of Graph Kernels on a Collaborative Recommendation Task. In *IEEE International Conference on Data Mining - ICDM*, pages 863–868. Citeseer.

Shih, Y.-K. and Parthasarathy, S. (2012). A single source k-shortest paths algorithm to infer regulatory pathways in a gene network. *Bioinformatics*, **28**(12), i49–i58.

Stojmirović, A. and Yu, Y.-K. (2012). Information flow in interaction networks II: channels, path lengths, and potentials. *Journal of computational biology : a journal of computational molecular cell biology*, **19**(4), 379–403.

Suthram, S., Beyer, A., Karp, R. M., Eldar, Y., and Ideker, T. (2008). eQED: an efficient method for interpreting eQTL associations using protein networks. *Molecular systems biology*, **4**(162), 162.

Tu, Z., Wang, L., Arbeitman, M. N., Chen, T., and Sun, F. (2006). An integrative approach for causal gene identification and gene regulatory pathway inference. *Bioinformatics*, **22**(14), e489–e496.

Voevodski, K., Teng, S.-H., and Xia, Y. (2009). Spectral affinity in protein networks. *BMC systems biology*, **3**(1), 112.

Zotenko, E., Mestre, J., O'Leary, D. P., and Przytycka, T. M. (2008). Why do hubs in the yeast protein interaction network tend to be essential: reexamining the connection between the network topology and essentiality. *PLoS computational biology*, **4**(8), e1000140.

Enhancer networks – Hidden layer of gene regulation

Justin Malin, Mohamed Radhouane Aniba, Sridhar Hannenhalli

Center for Bioinformatics and Computational Biology
Computational Biology, Bioinformatics, and Genomics Program
Department of Cell Biology and Molecular Genetics
University of Maryland, College Park, MD

Background

Eukaryotic transcription is intricately regulated at multiple levels, including epigenomic modifications, chromatin reorganization, and sequence-specific binding of TF to either proximal promoter regions or to distal enhancer/repressor regions of a gene [1]. Distal enhancers, which can regulate their target genes from long distances -- the most extreme case being the Shh gene's enhancer at ~1Mb away -- are especially important in regulating critical developmental and tissue-specific genes[2]. Recent advances in sequencing technologies have revealed putative distal enhancers based on various epigenomic marks, notably P300 binding [3]. Functionally linked genes tend to be co-expressed and are presumed to be co-regulated [4]. Gene networks based on co-expression patterns of gene pairs across multiple conditions and/or cell types reveal intricate organization of genes into pathways and functional groups [5]. Similar to functionally related genes, functionally related enhancers, i.e., those regulating functionally related genes, share TF binding sites and are likely to have spatio-temporal coordinated activity [6]. A network-level analysis of coordinated activities of distal enhancers has not been reported and such an analysis is likely to reveal higher order organization of a global transcriptional regulatory network mediated by distal enhancers. Analogous to using expression level to quantify transcriptional activity of a gene, DHS of an enhancer region has been proposed as a proxy for its condition-specific regulatory activity [7]. The ENCODE project has produced whole genome DHS profiles across numerous human cell types [8]. Analogous to using cross-condition expression correlation to infer gene networks, cross-condition DHS correlation can be used to infer enhancer networks. Using ~100K P300-bound regions as candidate enhancers, we have identified their correlated activity based on their DHS profiles across 72 human cell types, and followed with investigations of mechanisms and functional consequences of the correlated enhancer activity.

Methods Highlight

1. P300 bound regions in 4 cell types – HepG2, GM12878, H1-HESC and SK-N-SH_RA – were used as candidate enhancer regions, yielding 98,353 enhancers with average length of 500 bps;
2. We obtained DHS status (open or closed) for 72 tissue types in ENCODE yielding a 98,353 x 72 binary matrix. In order to minimize dependencies, tissues were clustered into 37 clusters yielding a 98,353 x 37 binary matrix.
3. Correlation between the activity of two enhancers was quantified using Mutual Information.
4. We controlled for cell type-specific DHS autocorrelation to detect significantly correlated enhancer pairs (Figure 1).

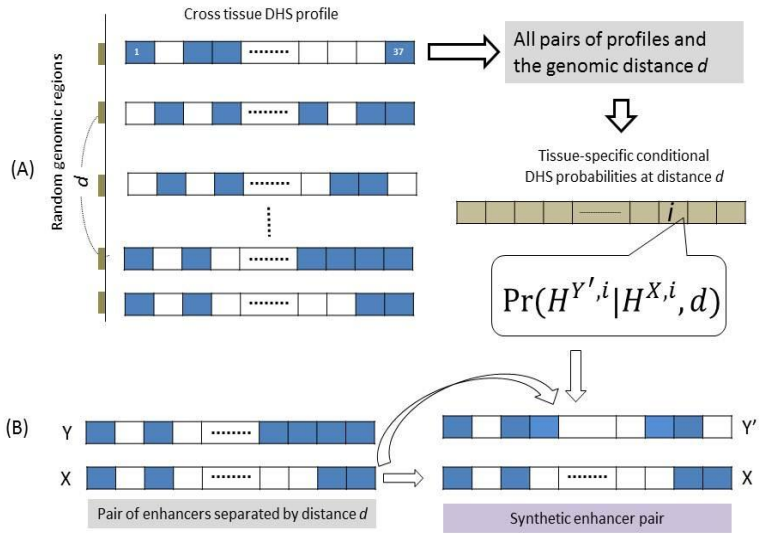


Figure 1. Generating the synthetic enhancer data to account for autocorrelation. (A) Starting with a large set of random genomic regions and their DHS profiles across 37 cell types, we estimate, for each cell type separately, the conditional probability of observing DHS at a location *Y'* given the DHS status at another location *X* at distance *d* from *X*. **(B)** Given a pair of enhancer DHS profiles (*X*,*Y*), we generate a synthetic pair of DHS profiles as (*X,Y'*) where *Y'* is randomly generated from *X* and the conditional probabilities estimated in (A). Blue: DHS=1 (open chromatin); white: DHS = 0 (closed chromatin)

Results Highlight

1. We exhaustively assessed ~35 million intra-chromosomal enhancer pairs separated by less than 12.5 Mb. Despite distance bin-specific FDR control, the fraction of enhancers that are significantly correlated declines with increasing distance (Figure 2). Across all bins, at an FDR of 1% we detect a total of 313,757 significant enhancer pairs, covering 32% of enhancers.

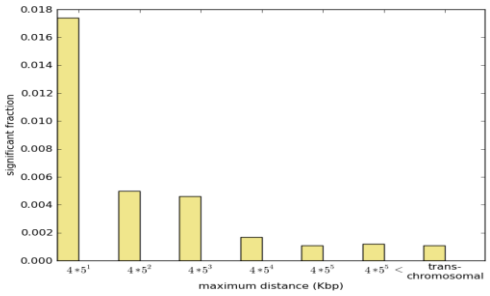


Figure 2. Chromatin states of a large number of enhancer pairs are significantly correlated. The plot shows the fraction of pairs with significant mutual information (*MI*) as a function of inter-enhancer distance. The plot is based on significant pairs after greedily removing pairs inducing transitive relationships.

2. Strong enhancers, those with higher expression levels of the nearest gene, tend to be correlated with fewer enhancers than weak enhancers but preferentially correlate with other strong enhancers, while weak enhancers are correlated with a greater number of enhancers and preferentially correlate with other weak enhancers.
3. Correlated enhancers tend to share common TF binding motifs. We identified 52 TF motifs significantly co-occurring in correlated enhancer pairs relative to uncorrelated enhancers. Using presence of shared motifs as features, correlated enhancers can be distinguished from uncorrelated ones with 73% accuracy. Several chromatin modification enzymes preferentially interact with these 52 TFs.
4. Using the gene closest to an enhancer as putative target, we found that the targets of correlated enhancers have correlated expression and are involved in common biological processes.

5. Correlated enhancers tend to be spatially close (although not highly significantly so) based on Hi-C data.
6. We constructed enhancer networks based on correlated activity and shared TF motifs, and found significant enrichment of specific biological processes among the putative gene targets of the enhancer modules (Figure 3).

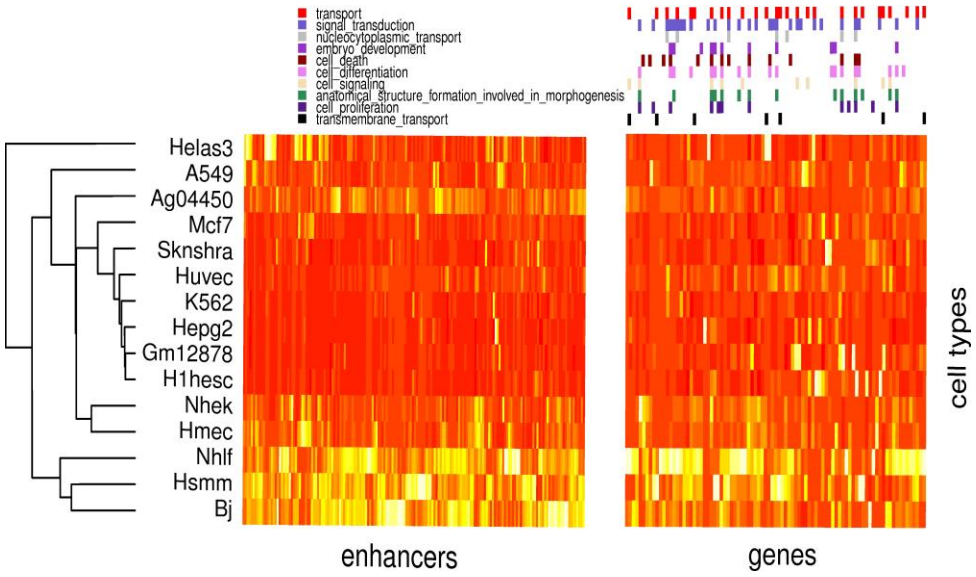


Figure 3: Tissue activity profile of an enhancer cluster and the corresponding target genes. Left: The tissue-specific DHS activity for 179 coordinately activated enhancers across 15 cell types. **Right:** Corresponding expression of the 98 target genes. The GO memberships for enriched terms each gene are shown above the heat plots.

Conclusions

Overall, our analysis suggests that functionally linked genes may be co-regulated by distal enhancers whose activities are regulated by common sets of TFs and mediated by both 3D chromatin structure as well as chromatin modification enzymes.

References

1. White, R.J., *Transcription by RNA polymerase III: more complex than we thought.* Nat Rev Genet, 2011. **12**(7): p. 459-63.
2. Lettice, L.A., et al., *A long-range Shh enhancer regulates expression in the developing limb and fin and is associated with preaxial polydactyly.* Hum Mol Genet, 2003. **12**(14): p. 1725-35.
3. Visel, A., et al., *ChIP-seq accurately predicts tissue-specific activity of enhancers.* Nature, 2009. **457**(7231): p. 854-8.
4. Stuart, J.M., et al., *A gene-coexpression network for global discovery of conserved genetic modules.* Science, 2003. **302**(5643): p. 249-55.
5. Dewey, F.E., et al., *Gene coexpression network topology of cardiac development, hypertrophy, and failure.* Circ Cardiovasc Genet, 2011. **4**(1): p. 26-35.
6. Narlikar, L., et al., *Genome-wide discovery of human heart enhancers.* Genome Res, 2010. **20**(3): p. 381-92.
7. Pique-Regi, R., et al., *Accurate inference of transcription factor binding from DNA sequence and chromatin accessibility data.* Genome Res, 2011. **21**(3): p. 447-55.
8. Dunham, I., et al., *An integrated encyclopedia of DNA elements in the human genome.* Nature, 2012. **489**(7414): p. 57-74.

Posters

1	Anna Lyubetskaya, Matthew Peterson, James Galagan	Reconstructing the regulatory network of TB: transcription factor binding distribution and properties *(see oral presentation for abstract)*
2	Meromit Singer, Idit Kosti, Lior Pachter, Yael Mandel-Gutfreund	Does intragenic DNA methylation determine differential exon expression? *(see oral presentation for abstract)*
3	Morgane Thomas-Chollier, Matthieu Defrance, Alejandra Medina-Rivera, Olivier Sand, Pierre Vincens, Carl Herrman, Sebastiaan H. Meijsing, Denis Thieffry, Jacques Van Helden	Deciphering genome-wide cis-regulation with RSAT: application to the glucocorticoid receptor
4	Rémy Nicolle, Mohamed Elati, Jennifer Southgate, and François Radvanyi	Modelling normal cells identifies master regulators in cancer
5	Joseph Wu, Beth Bragdon, Louis Gerstenfeld, Mayetri Gupta	Bayesian inference of gene regulatory networks from factorial time-course experiments with applications to bone fracture healing
6	Marcin P. Joachimiak, Cathy Tuglus, Mark van der Laan, Adam P. Arkin	Deep surveys of biological modules: K-biclustering gene expression and phenotype data
7	Alastair Kilpatrick, Stuart Aitken	Stochastic algorithms for motif discovery: a comparison of sampling strategies
8	Michael Dabrowski, Norbert Dojer, Izabella Krystkowiak, Bartek Wilczynski, Bozena Kaminska	Comparison of Jaspar, Transfac and Genomatix motif libraries on chip-seq data for 44 transcription factors
9	Izabella Krystkowiak, Michael Dabrowski	Integrating Nencki Genomics webservices via Taverna workbench
10	Limor Leibovich, Inbal Paz, Zohar Yakhini and Yael Mandel-Gutfreund	DRIMust: a web server for discovering rank imbalanced motifs using suffix trees
11	Jieun Jeong, Yuichi Nishi, Andrew P. McMahon	Polycomb repression and RNA polymerase in neural tube development
12	Alena van Bömmel, Mike Love, Ho-Ryun Chung, Martin Vingron	Detection of co-regulating transcription factors in 34 human cell types using predicted DNA-binding affinity on DNase hypersensitive sites
13	Marleen Claeys, Kathleen Marchal	Regulatory motif detection using different types of evolutionary conservation information
14	Joshua Welch, Jan Prins	Investigating the role of transcribed pseudogenes in breast cancer
15	Yaron Orenstein, Ron Shamir	Inferring binding site motifs from high-throughput in vitro data
16	Lex Overmars, Sacha A. F. T. van Hijum, Roland J. Siezen, Christof Francke	REPs, genetic insulators that enable differential regulation of gene expression in bacteria

17	Jocelyn Brayet, Remy Nicolle, Mohamed Elati	MiRnaBoost: Multi-view AdaBoost for microRNA target prediction
18	Stefan Naulaerts, Wim Vanden Berghe, Kris Laukens	Integrative biological itemset mining in cancer research
19	Galip Gurkan Yardimci, Gregory E. Crawford, Uwe Ohler	Prediction of genome-wide in vivo transcription factor binding using factor-specific DNase footprinting models

Deciphering genome-wide cis-regulation with RSAT: application to the glucocorticoid receptor

Morgane THOMAS-CHOLLIER[1], Matthieu DEFRANCE[2], Alejandra MEDINA-RIVERA[3], Olivier SAND[4], Pierre VINCENS[1], Carl HERRMAN[5], Sebastiaan H. MEIJSING[6], Denis THIEFFRY[1] and Jacques VAN HELDEN[5]

[1] Computational systems biology, Institute of Biology of ENS (IBENS), Paris, France
mthomas@biologie.ens.fr , thieffry@ens.fr , Pierre.Vincens@ens.fr

[2] Laboratory of Cancer Epigenetics, Faculty of Medicine, Université Libre de Bruxelles, Belgium
defrance@bigre.ulb.ac.be

[3] SickKids Research Institute, 101 College St. East Tower | Suite 15-306, Toronto, Ontario, Canada
alejandra.medina@sickkids.ca

[4] Génomique et maladies métaboliques, CNRS-UMR8199, Institut de Biologie de Lille, France
olivier.sand@good.ibl.fr

[5] Technological Advances for Genomics and Clinics (TAGC), INSERM U928 & Aix-Marseilles University, France
jacques.VAN-HELDEN@univ-amu.fr , carl.herrmann@univ-amu.fr

[6] Max Planck Institute for Molecular Genetics, Berlin, Germany
meijsing@molgen.mpg.de

Overview of RSAT

The regulatory sequence analysis tools (RSAT, http://rsat.ulb.ac.be/rsat/ and mirrors) is a software suite that integrates a large collection of modular tools for the detection of cis-regulatory elements in genomic sequences [1-3]. The web site has been running without interruption since 1998, and the suite has been continuously developed to accommodate novel types of data and experimental approaches over the years [2-4]. The suite includes programs for sequence retrieval, motif discovery, phylogenetic footprint detection, sequence scanning with regular expressions or position-specific scoring matrices, motif quality assessment and comparison, visualization and conversion utilities, along with a series of tools for random model generation and statistical evaluation. Genomes are regularly updated from various genome repositories (NCBI, Ensembl, UCSC browser) and the website currently supports 2517 genomes (March 2013).

RSAT enables genome-wide analysis of cis-regulatory elements with different types of input data: (i) groups of co-expressed genes produced by transcriptomic experiments, (ii) phylogenetically conserved regions and (iii) high-throughput binding data such as ChIP-seq.

In addition to motif discovery and pattern-matching approaches, the suite already provides various tools to analyse transcription factor binding motifs represented as matrices, including motif comparisons [3] and evaluation of matrix quality [5]. We are currently developing a motif clustering algorithm to ease the analysis of overlapping motifs (newly discovered or reported in databases) and assess potential motif diversity for a given transcription factor, in the context of motifs for cofactors.

The RSAT web server offers an intuitive interface, where each program can be accessed either separately or connected to the other tools. In addition, many tools are available as SOAP/WSDL web services, enabling their integration in programmatic workflows. Programs are documented with manual pages, while 'demo' buttons propose typical test cases. In addition, web tutorials and a series of published protocols help the users to master the different functionalities of RSAT [6-9], providing step-by-step guidelines about alternative options, as well as regarding the interpretation of results.

Cis-regulation analysis from high-throughput binding data

Several efficient and complementary motif discovery algorithms can predict transcription factor binding motifs from groups of co-expressed genes. Although these methods yield good results in yeast and bacteria

genomes, they are not suitable for vertebrates, due to the larger size and heterogeneity of non coding genomic sequences. The same algorithms nevertheless proved very efficient to analyse high-throughput transcription factor binding data, where the signal to noise ratio is higher. The workflow *peak-motifs* [10] was therefore developed to process large collections of peak sequences obtained from ChIP-seq or related technologies, to predict transcription factor binding motifs.

Most existing tools present limitations on sequence size, and they typically restrict motif discovery to a few hundred peaks, or to the central-most part of the peaks. To interpret genome-wide location data, there is a crucial need for time- and memory-efficient algorithms, interfaced as user-accessible tools to extract relevant information from high-throughput sequencing data.

Our workflow *peak-motifs* takes as input a set of peak sequences of interest, discovers key motifs, compares them with transcription factor binding motifs from various databases, predicts the location of binding sites within the peaks and exports them in a format suitable for visualization in the UCSC Genome Browser. Notably, all these steps, including motif discovery, are performed on the full-size sets of peak sequences, without restrictions on peak number or width.

The motif discovery step relies on a combination of algorithms that use complementary criteria to detect exceptional words (oligonucleotides and spaced motifs): global over-representation of oligonucleotides (*oligo-analysis*) or spaced pairs (*dyad-analysis*), heterogeneous positional distribution (*position-analysis*) and local over-representation (*local-word-analysis*).
The motif comparison step is performed by *compare-matrices* [3], which supports a wide range of scoring metrics and displays the results as multiple alignments of logos, enabling to grasp the similarities between a discovered motif and several known motifs. This feature is particularly valuable to reveal adjacent fragments of the discovered motif showing similarities with two distinct known motifs, suggesting a bipartite motif for two factors.
Sequences are scanned with the discovered motifs to locate binding sites, and their positioning within peaks is analyzed (coverage, positional distribution along peaks).
Peak-motifs generates an HTML report summarizing the main results and giving access to each separate result file. The report page includes links, allowing users to upload input peaks and predicted sites to the UCSC Genome Browser in order to visualize them in their genomic context.

We assessed the time efficiency of *peak-motifs* by analyzing data sets of increasing sizes (from 100 to 1 000 000 peaks of 100 bp each), with total sequence sizes ranging from 10 kb to 100 Mb. The computing time of the motif discovery algorithms integrated in *peak-motifs* increases linearly with sequence size and outperforms all the other existing motif discovery tools used in our comparison [10]. Data sets of several tens of megabytes are processed in a few minutes on a personal computer (the most efficient tool, *oligo-analysis*, treats 100Mb in 3min). This linear time response enables peak-motifs to scale up efficiently with sequence size, and allows us to provide an easy access via a web interface, without any data size restriction.

Current developments aim at extending *peak-motifs* to support other high-throughput data, including chromatin marks or DNaseI.

Condition-specific binding of GR

We extensively used RSAT to study the binding of the glucocorticoid receptor (GR), with the aim to better understand its specificity of action in various cell types. Glucocorticoids are steroid hormones that bind to its nuclear receptor (GR) that in turn recognizes specific DNA sequences to regulate the expression of target genes. The target genes of GR are highly cell-type specific and they mediate the various physiological effects of GR on processes including glucose metabolism (hepatocytes), anti-inflamatory effects (leucocytes) and increase of bone resorption, [11]. In particular, the anti-inflammatory effect of glucocorticoids are of enormous therapeutic importance and are widely used to treat a broad range of immune-related diseases e.g. allergies, asthma and arthritis. Unfortunately however, the beneficial therapeutical effects of glucocorticoids are accompanied by several side effects including muscle wasting,

metabolic changes and osteoporosis. The desired as well as the side effects of glucocorticoids are mediated by specific target genes that are regulated in a tissue specific manner. To better understand how glucocorticoids elicit highly tissue specific effects we used ChIP-seq to compare the genomic binding profile of GR in cell lines derived from B cells, bone and lung.

We made several observations when we compared the genomic binding profile of GR in these cell lines: first, there is little overlap of the bound regions, suggesting that genomic binding is highly cell-type specific. Moreover, the fraction of promoter-proximal GR binding differs greatly between cell-types. Our motif analysis suggests that GR recognizes similar binding sites (imperfect palindromes), but that the presence of sequence motifs for potential cofactors varies between cell lines. Preliminary experimental validation showed that genomic regions harbouring such cell-type specific cofactors could recapitulate cell-type specific regulation by GR arguing for their importance in directing cell-type specific GR actions.

Furthermore, we compared the binding specificity of two isoforms of GR that differ by a single amino acid which is a consequence of an alternative splice event. We found that the binding motif of these isoforms, GRalpha and GRgamma, differed slightly and functional luciferase reporter studies confirmed the importance of these differences in mediating isoform-specific transcriptional regulation. We do not know the exact reason for the difference between the binding motifs for GR alpha and gamma, but biochemical and structural studies indicated a role for isoform specific DNA binding affinity and conformation.

Current work on the GR involves the analysis of ChIP-exo datasets, with the goal to obtain more precise information on the mechanisms by which GR recognizes its binding sites.

References

[1] van Helden J. Regulatory sequence analysis tools. *Nucleic Acids Research* 31: 3593-6, 2003.

[2] Thomas-Chollier M, Sand O, Turatsinze JV, Janky R, Defrance M, Vervisch E., Brohee S & van Helden J. RSAT: regulatory sequence analysis tools. Nucleic Acids Res*earch* 36: W119-27, 2008.

[3] Thomas-Chollier M, Defrance M, Medina-Rivera A, Sand O, Herrmann C, Thieffry D, van Helden J. RSAT 2011: regulatory sequence analysis tools. *Nucleic Acids Research* 39: W86-91, 2011.

[4] Defrance M, & van Helden J. Info-gibbs: a motif discovery algorithm that directly optimizes information content during sampling. *Bioinformatics* 25: 2715-22, 2009.

[5] Medina-Rivera A, Abreu-Goodger C, Salgado-Osorio H, Collado-Vides J & van Helden J. Empirical and theoretical evaluation of transcription factor binding motifs. *Nucleic Acids Research* 39: 808–24, 2011.

[6] Turatsinze JV, Thomas-Chollier M, Defrance M & van Helden J. Using RSAT to scan genome sequences for transcription factor binding sites and cis-regulatory modules. *Nature Protocols* 3 : 1578-88, 2008.

[7] Defrance M, Janky R, Sand O.& van Helden J. Using RSAT oligo-analysis and dyad-analysis tools to discover regulatory signals in nucleic sequences. *Nature Protocols* 3 : 1589-603, 2008.

[8] Sand O, Thomas-Chollier M, Vervisch E & van Helden J. Analyzing multiple data sets by interconnecting RSAT programs via SOAP Web services-an example with ChIP-chip data. *Nature Protocols* 3 : 1604-15, 2008.

[9] Thomas-Chollier M, Darbo E, Herrmann C, Defrance M, Thieffry D & van Helden J. A complete workflow for the analysis of full-size ChIP-seq (and similar) data sets using peak-motifs. *Nature Protocols* 7: 1551-68, 2012.

[10] Thomas-Chollier M, Herrmann C, Defrance M, Sand O, Thieffry D, van Helden J.RSAT peak-motifs: motif analysis in full-size ChIP-seq datasets. *Nucleic Acids Research* 40: e31, 2012.

[11] Gross KL, Cidlowski JA: Tissue-specific glucocorticoid action: a family affair. Trends in endocrinology and metabolism: TEM 2008, 19(9):331-339.

Modelling normal cells identifies master regulators in cancer

Rémy Nicolle[1,2], Mohamed Elati[1], Jennifer Southgate[3], and François Radvanyi[2]

[1] iSSB, Institute of Systems and Synthetic Biology, CNRS FRE3561, University of Evry-Val-d'Essonne, Genopole Campus 1, Genavenir 6, 5 rue Henri Desbruères, 91030 Evry, France,
(remy.nicolle, mohamed.elati)@issb.genopole.fr,
[2] Institut Curie, CNRS UMR 144, 26 rue d'Ulm, 75248 Paris,cedex 05, France
francois.radvanyi@curie.fr
[3] Jack Birch Unit of Molecular Carcinogenesis, Department of Biology, University of York, United Kingdom.
js35@york.ac.uk

Motivation

Tumor cells display many functions possessed by their normal counterparts. Their ability to migrate, to proliferate, to attract new vessels and to exist in various differentiation states are properties also found in normal tissue during wound healing. Following tissue injury normal cells can operate these processes in a tightly regulated and coordinated manner leading to the healing of the wound.

We propose an original systems biology approach to identify and analyse the regulatory networks found in the normal states to then assess whether networks of the normal regenerative process are specifically maintained or altered in the tumor state. This strategy was applied to bladder cancer, a cancer derived from the bladder urothelium, because normal bladder urothelium can be grown in culture at various different stages of proliferation and differentiation thereby mimicking wound healing.

Constructing the network of normal proliferation and differentiation

Gene expression data of primary Normal Human Urothelium (NHU) non cancerous primary cell cultures in various states of differentiation and proliferation was considered as an *in vitro* model of wound healing and used to infer a large regulatory network. We applied LICORN[1], a data mining algorithm introduced by our team that infers the targets of transcription factors from genome wide expression data. LICORN was shown to be suitable for cooperative regulation and to scale up to the complexity of mammalian transcriptional networks. LICORN is able to find the set of Transciption Factors that cooperatively regulate the expression of a given gene. Furthermore, LICORN was previously applied[2] in yeast to infer a large regulatory network from gene expression data. The authors showed that the clusters of genes extracted from the inferred regulatory network had a higher functional enrichment than clusters based solely on gene expression.

Additionally to expression based information, the inferred normal regulatory network, comprising approximately 5000 genes and 400 co-regulators, was enriched with systematic promoter sequence analysis of known transcription factor binding sites model, public ChIP-chip and ChIP-Seq data as well as Protein Protein interactions between co-regulators.

Note that the rest of our approach is not dependent on the network inference method. Aside from the co-regulation information inferred by LICORN, any method able to infer large-scale regulatory networks such as ARACNE[3] or GENIE3[4] could be used.

Measuring context specific regulation activity

The concept of using the knowledge over the network structure was shown to be successful at identifying key regulators of specific phenotypes and processes [5, 6]. However, we were interested in a data transformation approach in which neither a predefined gene-signature nor sample classification was needed to identify central regulators.

In order to identify key regulators and to quantify their impact on their regulatory programs, we propose to measure the influence of regulator genes on their targets in a given sample. The idea is to be able to quantify the extent to which a Master Regulator (MR) is active on it's target genes in a given sample or set of samples. The measure is based on the divergence between the expression of the set of activated and repressed target genes of a given regulator in a given sample. The basic idea is that if a set of genes is effectively activated

and another repressed by the same MR, and that this MR is active, the activated set of genes should be over expressed and the repressed set should be under expressed. Therefore the more a MR is active on given sets of targets (activated and repressed) the greater the distance between these sets will be. Measuring this divergence will give an idea on the activity of a MR in a given sample, or set of samples, and more importantly on the pertinence of the structure of the network.

Interestingly, when measured for each regulator in each sample, the measure of regulatory influence produces a data set with the same number of samples but a reduced number of features representing the master regulators activity. Therefore, we proposed[7] to use this measure in the context of classification and feature extraction. We showed that the transformation of the data through the regulatory activity greatly improves the stability and robustness of models trained in different datasets.

Regulatory influences underline function of Master Regulators in normal and cancer cells

The regulatory network inferred from the NHU data pointed out several previously described regulator of normal urothelial differentiation and their validated gene targets. Alongside to these known MR, the computation of the influence characterized new MR as well as their involvement in normal urothelial differentiation, proliferation and growth arrest.

In an identical way the influence of the normal regulators was measured in 3 cohorts of 60, 120 and 180 bladder cancer transcriptomes (respectively from Stransky et. al. [8], the TCGA consortium and a private unpublished data set). In order to estimate to what extent the normal regulation of growth is conserved in tumor cells, the global influence (defined as the sum of squared influence in all samples for all MR) of the normal network was compared to the influence of 1000 randomly generated networks with similar topology (each regulator with the same number of target genes). The global influences were compared and shows that the normal network is significantly conserved, more influent, than any random network (150 times the standard deviation above the mean).

An analysis of the influence of normal regulators in bladder cancers pointed out a major loss of function of Master Regulators of urothelial differentiation. This loss of differentiation was also observed in the co-regulatory network in which known and novel regulators of normal differentiation form a dense network of cooperative regulators and are virtually all lost in most bladder tumors. Additionally to these results, several MR show the same activity profile in some bladder tumors than in the proliferating NHUs suggesting that the regulation driving normal proliferation is maintained during tumorigenesis.

References

1. M. Elati, P. Neuvial, M. Bolotin-Fukuhara, E. Barillot, F. Radvanyi, and C. Rouveirol. LICORN: learning cooperative regulation networks from gene expression data. *Bioinformatics*, 23(18):2407–2414, Sept. 2007.
2. E. Birmelé, M. Elati, C. Rouveirol, and C. Ambroise. Identification of functional modules based on transcriptional regulation structure. *BMC Proc*, 2 Suppl 4:S4, 2008.
3. A. A. Margolin, I. Nemenman, K. Basso, C. Wiggins, G. Stolovitzky, R. Favera, and A. Califano. ARACNE: An Algorithm for the Reconstruction of Gene Regulatory Networks in a Mammalian Cellular Context. *BMC Bioinformatics*, 7(Suppl 1):S7, 2006.
4. V. A. Huynh-Thu, A. Irrthum, L. Wehenkel, and P. Geurts. Inferring Regulatory Networks from Expression Data Using Tree-Based Methods. *PLoS ONE*, 5(9):e12776, Sept. 2010.
5. M. S. Carro, W. K. Lim, M. J. Alvarez, R. J. Bollo, X. Zhao, E. Y. Snyder, E. P. Sulman, S. L. Anne, F. Doetsch, H. Colman, A. Lasorella, K. Aldape, A. Califano, and A. Iavarone. The transcriptional network for mesenchymal transformation of brain tumours. *Nature*, 463(7279):318–325, Dec. 2009.
6. C. Lefebvre, P. Rajbhandari, M. J. Alvarez, P. Bandaru, W. K. Lim, M. Sato, K. Wang, P. Sumazin, M. Kustagi, B. C. Bisikirska, K. Basso, P. Beltrao, N. Krogan, J. Gautier, R. Dalla-Favera, and A. Califano. A human B-cell interactome identifies MYB and FOXM1 as master regulators of proliferation in germinal centers. *Molecular Systems Biology*, 6:1–10, June 2010.
7. R. Nicolle, M. Elati, and F. Radvanyi. Network transformation of gene expression for feature extraction. *11th International Conference on Machine Learning and Applications (ICMLA)*, pages 1–6, 2012.
8. N. Stransky, C. Vallot, F. Reyal, I. Bernard-Pierrot, S. G. D. de Medina, R. Segraves, Y. de Rycke, P. Elvin, A. Cassidy, C. Spraggon, A. Graham, J. Southgate, B. Asselain, Y. Allory, C. C. Abbou, D. G. Albertson, J. P. Thiery, D. K. Chopin, D. Pinkel, and F. Radvanyi. Regional copy number–independent deregulation of transcription in cancer. *Nat Genet*, 38(12):1386–1396, Nov. 2006.

Bayesian inference of gene regulatory networks from factorial time-course experiments with applications to bone fracture healing

Joseph Wu
Department of Biostatistics, Boston University School of Public Health
Email: josephwu@bu.edu

Beth Bragdon
Department of Orthopedic Surgery, Boston University School of Medicine
Email: bragdon@bu.edu

Louis Gerstenfeld
Department of Orthopedic Surgery, Boston University School of Medicine
Email: lgersten@bu.edu

and

Mayetri Gupta
School of Mathematics and Statistics, University of Glasgow
Email: mayetri.gupta@glasgow.ac.uk

Abstract

Bone fracture healing recapitulates many aspects of embryonic skeletal development. Besides age, metabolic conditions, and the presence of pharmacological agents, gender and genetic predisposition may affect the cellular environment and skeletal repair processes. Also, the fracture repair process takes place in stages over a long period of time with different networks of genes involved at different times. To improve the quality and speed of the repair process, it is important to understand how the genes involved behave under critical experimental conditions and longitudinally over time. As experimental designs become more complex such as in factorial time-course microarray studies, it becomes more challenging to answer questions of interest, such as how two experimental factors interact in their effects on gene expression over time. One may want to detect possible interactions between an experimental treatment and another factor while concurrently grouping genes showing similar effects into clusters. Only a few existing methods are able to simultaneously infer differential expression and take gene clustering into account. Motivated by the need to fully model a factorial time-course gene expression experiment, we propose a novel Bayesian statistical approach that can simultaneously estimate the longitudinal model signals under a factorial design and assign genes into biologically meaningful clusters, using fast hybrid MCMC algorithms. A unique feature of our framework is that all information about gene expression can be interpreted at all three levels– longitudinal, factorial, and transcriptional.

Deep Surveys of Biological Modules: K-biclustering Gene Expression and Phenotype Data

Marcin P. Joachimiak[1*], Cathy Tuglus[2], Mark van der Laan[2], Adam P. Arkin[1,3]
[1] Physical Biosciences Division, Lawrence Berkeley National Laboratory, Berkeley, CA, 94720, USA
[2] Dept. of Biostatistics, University of California, Berkeley, 94720, USA
[3] Dept. of Bioengineering, University of California, Berkeley, 94720, USA
*To whom correspondence should be addressed: MJoachimiak@lbl.gov

Abstract:

Experimental efforts are targeting genomes, cells, and populations of organisms with a widening array of high throughput experimental techniques. In light of this bountiful data it is advantageous to: a) simultaneously query multiple data types to uncover new biological associations, b) jointly determine confidence across data types, and c) systematically form hypothesis from multiple lines of evidence. We have developed an accurate and sensitive biclustering algorithm, Massive Associative K-biclustering (MAK), for the discovery of biological data associations across multiple data types. The algorithm framework models data archetypes, such as object-by-value (e.g. gene expression), object-by-feature (e.g. phylogenetic profiles), and object-by-object (e.g. protein interactions). The objects can be for example genes, proteins, regulators, orthologous sequence families, or experiments. For each data archetype we designed statistical criteria to detect the expected association patterns. True associations in biological data are mostly unknown, therefore to evaluate biclustering methods we design a simulated dataset modeled on yeast gene expression data, with implanted associations forming known patterns. Using this and other evaluations we find that MAK compares favorably to other biclustering methods.

We applied the MAK algorithm to reconstruction of a condition-specific transcriptional co-expression network for Saccharomyces cerevisiae using combinations of gene expression, experimentally determined transcription factor binding, protein interaction, and phylogenetic profile data. Using gene expression data alone we find more biclusters with higher enrichment for known transcription factor regulation and functional terms than other biclustering methods. Pooling biclusters independently discovered using different data type combinations leads to an improvement in most evaluation measures. We also used the discovered biclusters to generate an annotated co-expressed module network, integrating the MAK biclusters, with human-curated groups of experimental conditions, human-curated functional terms, categories, and cellular localization, known regulation, and known sequence motifs. We were able to assign putative functions and regulation to a number of novel biclusters.

Stochastic algorithms for motif discovery: a comparison of sampling strategies

Alastair M. Kilpatrick and Stuart Aitken*
School of Informatics, University of Edinburgh

April 26, 2013

TFBS motifs are short DNA sequence patterns that have important roles in gene transcription and regulation. Discovery of these sequences remains an important task in the wider challenge of understanding the mechanisms of gene expression; consequently, there is much continuing interest in developing algorithms to computationally discover TFBS motifs.

The EM algorithm [4] is the basis of a number of algorithms for motif discovery (most notably the popular MEME algorithm [1]). However, it suffers from several well-known limitations: it is strongly dependent on its initial position and can converge to a saddle point of the likelihood function rather than a local maximum. A stochastic version of the EM algorithm has been shown to alleviate these limitations in theory [3] and has been implemented in a motif discovery context by Bi (using the OOPS, or One Occurrence Per Sequence model) as the SEAM algorithm [2]. In this study we compare a Metropolis independence sampler with the roulette wheel selection used in SEAM in order to evaluate the potential performance benefits and computational cost: the motivation for this study is to determine if it is possible to reduce the running time of the algorithm by designing a strategy where samples could be drawn from an input sequence without having to evaluate the probability of each possible motif start site being an occurrence of the motif. The correctness of the recovered motifs is assessed using the standard measures of site-level sensitivity (sSn) and positive predictive value ($sPPV$).

Background

The idea underlying SEAM is to replace the computation and maximisation of the expected complete-data log likelihood function by the much simpler estimation of the posterior distribution for each input sequence, simulating a 'pseudo-sample' from this distribution and updating the model parameters based on the pseudo-complete samples [2]. This method is equivalent to the weighted 'roulette wheel selection' (sometimes known as 'fitness proportionate selection') method in genetic algorithms. Having sampled each input sequence, a proposal model is constructed from the samples and the current model updated to the proposal model if the Metropolis ratio is satisfied [2].

Using this method, the probability that a given position j in input sequence i is a motif occurrence ($Z_{i,j}$) must be enumerated for every position in i at every EM iteration in order to calculate the density. This requires considerable computation and may be inefficient, especially at later EM iterations when the majority of $Z_{i,j}$ values are expected to be near zero. This motivates the current study: is it possible to sample from an input sequence without having to evaluate $Z_{i,j}$ at every position? One potential solution is to use Markov Chain Monte Carlo (MCMC) to sample from our input sequence.

Method and Results

The simplest MCMC strategy (Metropolis algorithm) uses an independence sampler as the proposal distribution; this simplifies the calculation of the acceptance probability. Clearly, this method is only an improvement on the roulette wheel selection method if the cost of drawing k samples is substantially smaller than the cost of evaluating $Z_{i,j}$ at every possible motif start site. It is well known that the

*AMK is supported by an EPSRC Doctoral Studentship. AMK and SA are funded by BBSRC grant BB/I023461/1 (Bayesian evidence analysis tools for systems biology).

Metropolis algorithm with independence sampler can be shown to converge to a target distribution when this distribution is well-behaved. While analysis of the posterior distribution for a given input sequence shows that this distribution is not well-behaved at all, we have shown that this general result holds true in the context of motif discovery for large k.

A modified version of SEAM was implemented, replacing the roulette wheel selection method with the Metropolis independence sampler for each input sequence. The Metropolis independence sampler was implemented within SEAM, replacing the roulette wheel selection method for each input sequence. Overall performance was assessed by running the modified SEAM algorithm with 1,000 random seeds, choosing the best result based on the motif energy function provided by Bi and calculating the site-level sensitivity (sSn) and site-level positive predictive value ($sPPV$) for the corresponding motif model. Bi's motif energy function is related to the sequence binding or structural configuration free energy, widely used in motif discovery algorithms [2].

Both the original roulette wheel and modified SEAM algorithms were tested on a small collection of datasets containing previously characterised *E. coli* TFBS motifs extracted from the RegulonDB database. Initial tests with $k = 1,000$ (around five times the number of possible motif start sites) returned similar results to the roulette wheel selection method, showing the Metropolis independence sampler converging to the target distribution. In some cases, the maximum value of the motif energy function was increased when using the independence sampler (i.e. the output motif model was stronger), giving a corresponding improvement in sSn and $sPPV$. While this improvement is encouraging, the main disadvantage of this result is that drawing 1,000 samples from each input sequence takes substantially longer than simply enumerating every position and drawing a sample from the roulette wheel. It is clear that the next step is to investigate whether this trend continues when k is decreased.

In tests on a single input sequence, the Metropolis independence sampler with smaller k shows relatively poor convergence. However, it may still give reasonable results when applied to the SEAM algorithm, as SEAM takes a sample from each input sequence and takes the consensus of all samples in order to form a new proposal model. It follows that even if the chosen sample for a single input sequence is relatively poor, this may be alleviated by the chosen samples from other input sequences. It is possible that the independence sampler still allows the stochastic EM algorithm at the heart of SEAM to converge, albeit at a slower rate than before.

Further tests were carried out with $k = 200$ (i.e. around the number of possible motif start sites) and $k = 20$ (i.e. around 0.1 of the number of possible motif start sites). In addition, the number of EM iterations was varied in order to determine whether increasing this would improve situations with fewer MC samples. The results of these tests show that, overall, as k is reduced, the maximum value of the motif energy function decreases (i.e. the output motif model becomes weaker), often reducing sSn and $sPPV$ as the number of true positive site predictions decreases.

Table 1 illustrates some of the results of the comparison of sampling strategies. In the case of the Ada motif, the Metropolis independence sampler improves the sSn and $sPPV$ of a motif which was not discovered well by the roulette wheel sampling method. This test also illustrates a slight increase in motif energy; this increase is also noted in other datasets. In the case of the MetR motif, while the sSn and $sPPV$ results for the Metropolis method with large k match those for the roulette wheel sampling method, this performance decreases as k is decreased. In both cases, as k decreases, the maximum motif energy also decreases. Our results also show that for k greater than the number of possible motif start sites, increasing the number of EM iterations may slightly increase the motif energy of the result. However, for small k, the overall result is poor and increasing the number of EM iterations makes little difference to the maximum motif energy (there is very little improvement over randomly choosing motif positions within the dataset). While increasing the number of EM iterations may lead to a small improvement in the mean motif energy over 1,000 random seeds, this improvement is not enough to offset the effect of reducing k.

Conclusions

Although the Metropolis algorithm with independence sampler is a relatively simple sampling strategy, this approach is shown to give surprisingly good recovery of motifs based on site-level sensitivity and

Motif	Ada				MetR			
Method	Roulette	Metropolis			Roulette	Metropolis		
MC samples	-	1000	200	20	-	1000	200	20
EM iterations	-	500	1500	2000	-	500	1500	2000
sSn	0.00	0.25	0.25	0.25	0.71	0.71	0.14	0.00
$sPPV$	0.00	0.25	0.25	0.25	0.71	0.71	0.14	0.00
Motif energy	-24.03	-23.17	-30.02	-45.80	-43.94	-47.52	-72.23	-84.72

Table 1: Results of sampling strategy comparison for two *E. coli* TFBS motifs.

positive predictive value. Implementing this approach and using large numbers of Monte Carlo samples is also shown to often return stronger motif models, based on Bi's motif energy function. We note the high computational cost of drawing large numbers of samples using the independence sampler, however its performance in this study indicates the potential in exploring alternative sampling strategies as replacements for the roulette wheel method.

References

[1] T. L. Bailey and C. Elkan. Unsupervised learning of multiple motifs in biopolymers using Expectation Maximization. *Machine Learning*, 21:51–80, 1995.

[2] C. Bi. SEAM: a stochastic EM-type algorithm for motif-finding in biopolymer sequences. *Journal of Bioinformatics and Computational Biology*, 5(1):47–77, 2007.

[3] G. Celeux, D. Chauveau, and J. Diebolt. On stochastic versions of the EM algorithm. *Rapport de Recherche-Institut National de Recherche en Informatique et en Automatique*, 1995.

[4] A. Dempster and N. Laird. Maximum likelihood from incomplete data via the EM algorithm. *Journal of the Royal Statistical Society. Series B (Methodological)*, 39(1):1–38, 1977.

Comparison of Jaspar, Transfac and Genomatix motif libraries on chip-seq data for 44 transcription factors

Michal Dabrowski[1]*, Norbert Dojer[2], Izabella Krystkowiak[1], Bartek Wilczynski[2], Bozena Kaminska[1]

[1] Nencki Institute of Experimental Biology, Warsaw, Poland
[2] Faculty of Mathematics Informatics and Mechanics, University of Warsaw, Poland

*presenting author, e-mail: m.dabrowski@nencki.gov.pl

For vertebrates, there are three major collections of TFBS motifs: public Jaspar and commercial Transfac and Genomatix. We compared performance of the three libraries, in terms of coverage, specificity, and sensitivity, using the chip-seq data for 44 transcription factors (TFs) as the positive sets, and third exons as the negative set. Each commercial library was used with its supplied scanner (Match and MatInspector), and all the three libraries were used with the same two open source scanners (Bio.Motif and matrix-scan).

The coverage (number of represented TFs) was highest for Genomatix (37), followed by Transfac (33), and by Jaspar (21). The average specificity and sensitivity was practically identical for all three libraries, when used with the same scanner. The two open-source scanners outperformed the two commercial scanners, by resulting in higher average sensitivity for the same average specificity. The use of Genomatix matrix families busted sensitivity, at the cost of a drop in specificity.

With Bio.Motif scanner, we analyzed the full ROC curves for all the motifs from the three libraries. We investigatied utility of different ways of parametrization of the ROC curves to automatically set the thresholds in a way that maximizes the balanced accuracy (average of specificity and the sensitivity). Our results demonstrate that the optimal value of the threshold is dependent on the information content of the motif.

Integrating Nencki Genomics webservices via Taverna workbench

Izabella Krystkowiak, Michal Dabrowski*
Nencki Institute of Exeperimental Biology, Warsaw, Poland.
**presenting author, e-mail: m.dabrowski@nencki.gov.pl*

Under the Nencki Genomics project (http://www.nencki-genomics.org), we developed a set of webservices for analysis of gene co-expression, cis-regulatory regions, and functional annotations; on the basis of user-supplied genomic or expression data and the large body of public regulatory genomic data provided by the Nencki Genomics Database [1]. The webservices use the well-defined SOAP/WSDL interface and are divided into two sets: genomic and expression. The genomic webservices provide functionalities of mapping regulatory areas to genes, intersecting regulatory areas, and intersecting areas with known TFBS motifs (both Jaspar and Transfac), identified genome-wide. Notably, we provide a webservice function, which plots a graphical representation of selected NGD content in the flank of transcription start site of a chosen gene. The expression webservices can be chained to provide a typical workflow of analysis of transcriptomic data, from pre-processed gene expression data (probes/genes x conditions), through probesets mapping and data transformation, to identification of differentially expressed genes, clustering and visualization. At each step of the analysis, the results can be returned to the user as a TSV file, piped to the next step, or stored in the underlying databases (providing access right control) for future use, sharing the data with others, or making them public.

Taverna Workbench (http://www.taverna.org.uk/) is a rapidly developing open-source workflow management system, which we use for the integration of the Nencki Genomics webservices. Taverna's graphical user interface (GUI) makes these functionalities accessible to a broad biological community of users.

[1] Krystkowiak et al. Nencki Genomics Database – Ensembl funcgen enhanced with genome-wide TFBS motifs, intersections and user data. *Database*, under revision.

DRIMust: a web server for Discovering Rank Imbalanced Motifs Using Suffix Trees

Limor Leibovich[1*], **Inbal Paz**[2*], Zohar Yakhini[1,3] and Yael Mandel-Gutfreund[2]

[1]Department of Computer Science, Technion - Israel Institute of Technology, Technion City, Haifa 32000, Israel

[2]Department of Biology, Technion - Israel Institute of Technology, Technion City, Haifa 32000, Israel

[3]Agilent Laboratories Israel, 94 Em Hamoshavot Road, 49527 Petach-Tikva, Israel

Abstract

Cellular regulation mechanisms that involve proteins and other active molecules interacting with specific targets often involve the recognition of short sequence elements. Studies that focus on measuring and investigating sequence based recognition processes make use of statistical and computational tools that support the identification and understanding of sequence motifs. We present a new web application, named DRIMust, freely accessible through the website: http://drimust.technion.ac.il, for de-novo motif discovery services. The DRIMust algorithm is based on the minimum-hypergeometric (mHG) statistical framework using suffix trees for an efficient enumeration of motif candidates. DRIMust takes as input ranked lists of sequences in FASTA format and returns motifs that are over-represented at the top of the list, where the determination of the threshold that defines top is data driven. The resulting motifs are presented individually with an accurate *p*-value indication and as a Position Specific Scoring Matrix (PSSM). Comparing DRIMust to other state-of-the-art tools demonstrated significant advantage to DRIMust both in result accuracy and in short running times. Overall, DRIMust is unique in combining efficient search on large ranked lists with rigorous *p*-value estimation for the detected motifs.

Polycomb repression and RNA polymerase in neural tube development

Jieun Jeong[1], Yuichi Nishi[2], Andrew P. McMahon[2]
[1]University of Pennsylvania School of Medicine, Philadelphia, PA [2]Keck School of Medicine, USC, L.A., CA

I. INTRODUCTION

Embryo development involves differentiation of tissues that starts from embryonic stem cells. The emerging tissues send and receive signaling molecules and in response differentiate further, and the proper timing and spatial characteristics of each stage are important for correct development. This requires precise control of the expression of key transcription factors that in turn control gene expression in their respective cell types. Here we establish new aspects of activity of polycomb repression complexes, PRC, which are involved in this process. Activity of PRC complexes forms additional layer of control for genes that are activated/deactivated with making/removing epigenetic modifications at promoter regions such as H3K4me3 and H3K27ac. It was shown previously that PRC-2 complex induces tri-methylation of lysine 27 of histone H3, epigenetic mark H3K27me3, and this modification recruits PRC-1 complex which makes further epigenetic changes that may prevent the conversion of RNA polymerase II (pol2) to a conformation that produces gene transcripts. This prevents gene expression. Alternatively, PRC histone modifications induce pol2 form that does produce gene transcripts but orders of magnitude less efficiently than the standard productive form. It was also shown previously that PRC activity is more frequent in early cell development where it affects 20-25% of genes, and that most genes repressed by PRC are eventually expressed.

II. PURPOSE AND HYPOTHESIS

For most genes the expression level is determined by the cell without PRC activity by regulating the activation through H3K4me3, H3K27ac etc., which opens the question of the benefit of an extra level of regulation. We show that PRC allows to recruit pol2 to genes without expressing them. When PRC activity is being reduced, the expression of gene increases as pol2 changes to productive forms. The speed of that increase depends on the amount of pre-loaded pol2. The differences in that speed for different genes are particularly important when two master transcription factors are simultaneously stimulated and they repress each other, and in the stable state only one of two factors is expressed. In the development of the neural tube, this situation is present for Nkx2-2 and Olig2, the key factors of two adjacent layers in the neural tube, pV3 and pMN (motor neurons), and possibly with other boundaries between the layers.

III. CONCLUSIONS

In the development of neural tube, neural tube cells initially represent EB type with high levels of Oct4. With a certain combination of signals, EB cells eliminate Oct4, activate Sox2 and stimulate expression of Pax6. Then Shh signal from the notochord creates Pax6-free layers on the ventral side, FP, pV3, pMN, pV2

Qualitative narrative of Balaskas et al. At the beginning of the EB to NEB transition, three key factors (Pax6, Olig2, Nkx2.2) have low levels. Pax6 increases first, activated by RA. Next, in the ventral layers Olig2 increases in response to Shh ➔ Gli1, while Nkx2-2 is repressed by Olig2 and Pax6. Olig2 represses Pax6 which decreases. However, Olig2 is a less effective repressor of Nkx2-2 than Pax6. When the net impact of Gli1 (activator from Shh pathway), Oli2 and Pax6 (repressor) is sufficient to stimulate Nkx2-2, the latter responds rapidly, thus eliminating Pax6 and Olig2 from pV3 (but not from pMN)

This system relies on combinations of activating impact of Gli1 and repressors and we hypothesize that these signals independently regulate pol2 recruitment and PRC activity. Redundant mechanisms of gene control levels enable different types of responses to simultaneous activation and repression and precise control of patterns formed in the tissue. We observe that Nkx2-2 has the highest levels of pol2/H3K36me3 (ratio), Pax6 is intermediate and Olig2 has the lowest. We conjecture that this determines the speed of reaction of these genes to activating signals as needed by dynamic control of gene expression.

Alena van Bömmel[1], *Mike Love*[1], *Ho-Ryun Chung*[2] *and Martin Vingron*[1]

(1) Computational Molecular Biology, Max Planck Institute for Molecular Genetics, Berlin, Germany
(2) Otto-Warburg Laboratory, Computational Epigenomics Group, Max Planck Institute for Molecular Genetics, Berlin, Germany

Detection of co-regulating transcription factors in 34 human cell types using predicted DNA-binding affinity on DNase hypersensitive sites.

BACKGROUND: Cell-type-specific gene expression is regulated by combinatorial interactions among transcription factors (TFs) binding to the DNA. Information about TFs' binding affinity to distal and proximal regulatory sequences can help determine which combinations of factors work together to regulate their target genes in cell-type-specific manner.

RESULTS: In this study, we provide detection of co-regulating TF pairs in 34 healthy human cell types which is based on statistical analysis of estimated ranked lists of TFs' target regions. Specifically, we first scanned all cell-type-specific DNase hypersensitive sites (DHSs) for single TF-DNA binding affinities using known motifs for 160 TFs and ranked the DHSs by their predicted binding affinity separately for each TF. We then studied the similarity of pairs of the ranked lists stratified by cell type by applying a statistical test for multiway contingency tables. Our significant TF pairs defined by the test in each cell type were validated by known protein-protein interactions (PPIs) and by detected co-binding of TFs in ChIP-seq data. We found that the known PPIs are significantly enriched (up to 12 fold) in the groups of our predicted co-regulating TFs and that we can recover a majority (56%) of predicted co-binding TF pairs from the ChIP-seq analysis. Furthermore, the predicted co-regulating TFs are supported in literature to be active regulators in the corresponding cell types.

CONCLUSION: Our findings show that the cell-type-specific gene expression is regulated by a large number of combinatorial TF interactions with dominating central regulators. However, the TF interaction networks substantially differ even for related cell lines.

RegGen SIG 2013 - Abstract for poster presentation

Authors:
Marleen Claeys [1], Kathleen Marchal [1.2.3]
[1] CMPG-bioi, Department of Microbial and Molecular Systems, KU Leuven, Belgium, [2] Department of Plant Biotechnology and BioInformatics, [3] VIB Department of Plant Systems Biology, UGhent, Belgium

Title:
Regulatory motif detection using different types of evolutionary conservation information.

Abstract (250 words at most):
Computational methods, which search *de novo* for conserved sites in a non-functional background, have been proven successful for the prediction of regulatory motifs. Conservation is typically quantified by overrepresentation (in the promotor regions of coregulated genes) and/or by evolutionary conservation in the promotor regions of orthologous genes.
We present 3 different adaptations of our well known overrepresentation based motif detection tool MotifSampler in order to search in both spaces of conservation simultaneously. Each adaptation quantifies evolutionary conservation in a different way : from 1) robust counting (MotifSampler-cPSP, allows the use of a position specific prior built by counting the occurrences of sites in orthologs), over 2) a comparative approach (NOrthoMotifSampler, assumes motif evolution is 'slower than' background evolution in orthologs), to 3) explicit evolution modeling (PhyloMotifSampler, uses a motif evolution model adapted from FelsensteinF81). MotifSampler-cPSP is most suitable for datasets with phylogenetically closely related orthologs whereas NOrthoMotifSampler is most sensitive in detecting sites that have mutated over time (in distantly related orthologs). PhyloMotifSampler finds motifs well in closely as well as distantly related orthologs yet with a lower site annotation accuracy and longer runtime. Neither of the newly developed tools requires prealignment of the orthologs which makes them attractive for datasets where such alignment is unreliable.

Investigating the Role of Transcribed Pseudogenes in Breast Cancer

Joshua Welch[1], Jan Prins[1]

[1]Department of Computer Science, The University of North Carolina at Chapel Hill, Chapel Hill, NC, USA

Abstract

Pseudogenes are genomic sequences closely resembling genes but possessing sequence differences that prevent them from encoding functional proteins. Although the human genome contains thousands of pseudogenes, these sequences are generally disregarded in functional genomic studies and are widely viewed as non-functional. However, there is increasing evidence that some pseudogenes are actually transcribed into RNA and can contribute to cancer when dysregulated. In particular, pseudogene transcripts can sequester miRNAs that would otherwise target mRNAs. In this role pseudogenes function as competing endogenous RNA (ceRNA).

To investigate the hypothesis that transcribed pseudogenes play a role in cancer, we developed a bioinformatics method for studying pseudogene transcription using RNA-seq and applied this method to 820 breast cancer samples from The Cancer Genome Atlas project. We incorporated sample-paired gene and miRNA expression data and miRNA target prediction to assess the potential ceRNA function of transcribed pseudogenes. We also performed a clustering analysis using the pseudogene expression data, determining how variation in pseudogene expression relates to known breast cancer subtypes.

Our results indicate that many pseudogenes are transcribed in breast cancer. A subset of these exhibit significant differential expression between tumor and normal samples. The expression levels of the differentially expressed pseudogenes correlate with a number of known cancer-related genes. Furthermore, our analysis incorporating miRNA target prediction and miRNA expression data suggests that a number of transcribed pseudogenes are strong candidates for ceRNA function. Taken together, these results indicate that pseudogene transcription in cancer plays a larger role than previously appreciated.

INFERRING BINDING SITE MOTIFS FROM HIGH-THROUGHPUT IN VITRO DATA

Yaron Orenstein[1], Ron Shamir[1]
[1] Blavatnik School of Computer Science, Tel-Aviv University, Tel-Aviv, Israel

Understanding gene regulation is a key challenge in today's biology. The new technologies of protein binding microarrays (PBMs) and high-throughput SELEX (HT-SELEX) allow measurement of the binding intensities of one transcription factor (TF) to an enormous number of synthetic double-stranded DNA probes in a single experiment. The PBM technology is based on microarrays, while HT-SELEX uses deep sequencing. The ChIP-seq technique uses deep sequencing to identify bound DNA segments *in vivo*. A key computational challenge is inferring the binding site motif of the tested TF from the experimental data.

Recently, a new study (Jolma *et al.* Cell 2013) reported the results of hundreds of HT-SELEX experiments on human TFs, including many TFs covered by PBM technology. We assessed the similarities and differences between PBM and HT-SELEX technologies, and measured the performance of binding models produced by each technology in predicting *in vivo* binding. Using published HT-SELEX-derived models to predict PBM bound probes results in worse performance than PBM-derived models (average AUC 0.78 compared to 0.89). Average correlation between the top k-mers ranked by the two technologies is just over 0.5. HT-SELEX-derived models are slightly better in predicting *in* vivo binding (average AUC 0.72 compared to 0.7 on ChIP-seq data).

Our analysis currently focuses on measuring and correcting for biases. We observed GC-bias in the sequencing files, as well as systematic enrichment of specific k-mers. We will report progress towards the development of a robust computational pipeline to generate an accurate binding model from HT-SELEX data.

REPs, genetic insulators that enable differential regulation of gene expression in bacteria

Lex Overmars[1,2] , Sacha A. F. T. van Hijum[1,2], Roland J. Siezen[1,2] and Christof Francke[1,2]

[1)]Radboud University Medical Centre, Centre for Molecular and Biomolecular Informatics, Nijmegen, The Netherlands [2)]Netherlands Bioinformatics Centre, 260 NBIC, PO Box 9101, 6500HB Nijmegen, The Netherlands [3)]TI Food and Nutrition, P.O. Box 557, 6700 AN Wageningen, The Netherlands, [4)]NIZO food research, P.O. Box 20, 6710 BA Ede, The Netherlands, [5)]Kluyver Centre for Genomics of Industrial Fermentation, P.O. Box 5057, 2600 GA Delft, The Netherlands

Background: Repetitive Extragenic Palindromic elements (REPs) are short palindromic sequences, commonly found in enteric bacteria. REPs (i) are almost exclusively found in the intergenic space, often arranged in repeats (BIMEs); (ii) occur in high abundance and; (iii) are highly conserved within a genome. Various biological roles have been proposed, however none of them provides a common functional denominator. We therefore decided to investigate the commonality using a comparative genomics approach.

Results: *E.coli* REPs were identified using the related 29bp conserved sequence. We observed a biased distribution of REPs with respect to the ORFs: REPs are not found between divergent gene-pairs and predominantly located between convergent gene-pairs. A set of 465 publicly available microarrays (M3D) was used to explore the effects of REPs on transcription under various conditions. This analysis revealed an association between REP-related gene-pairs and higher expression levels. This association is also evident when Codon Adaptation Index values were compared. We identified microarrays with significant effects on gene-REP-gene pair (co-) expression. These arrays all represented the transcriptional response to certain kinds of stress such as biofilm formation and aerobiosis.

Conclusions: This study shows that REPs potentially have a global role in regulation of differential expression. Our results imply that REPs enable differential expression specifically in cases were transcription-driven DNA supercoiling can arise, i.e. expression of convergent gene-pairs and transcription regulated by an alternative promoter. Our findings suggest that the phenomenon of REP-enabled differential expression is linked to the bacterial stress response in *E. coli*.

MiRnaBoost: Multi-view AdaBoost for microRNA target prediction

Jocelyn Brayet, Remy Nicolle, Mohamed Elati
institute of Systems and Synthetic Biology
Genopole Campus 1 - Genavenir 6
5 rue Henri Desbruères - F-91030 EVRY cedex
mohamed.elati@issb.genopole.fr

MicroRNAs are short (21-25 nt) non-coding RNAs that repress the expression of their direct targets (Bartel, 2009). Building an accurate binding model for a microRNA is essential to differentiate its true binding targets from spurious ones (Khorshid 2013). So far, conventional approaches to prediction of microRNA binding sites have all relied on local sequence information only, in a way or another. In this work we devise a novel machine learning system, MiRnaBoost, to build a microRNA binding classifier by combining sequence, expression and position information-based classifiers. Currently, sequence-based prediction methods are not fully capturing microRNA target preferences, nor context specific regulations. To overcome the limitation of sequence-only miRNA-gene interaction prediction, MiRnaBoost complements a sequence based classifier (miRanda) with two additional supervised models trained on different views i) the expression levels of both the miRNA and the target gene (Huang 2007), ii) the pattern of the genomic position (Elati 2013) of the targets of a miRNA. MiRnaBoost combines these weak classifiers using a modified version of the Adaboost algorithm, which manages to combine and improve together classifiers trained on the same instances but on different views (Zhijie 2010; Elati 2013).
Based on cross-validation analysis over the microRNAs with the most validated targets in TarBase, MiRnaBoost consistently outperforms conventional methods exploiting only sequence information. The main advantage of MiRnaBoost is that it lowers the false positive rate. Furthermore, MiRnaBoost predicted miRNA target sets are more consistently annotated with GO terms than similar sized random subsets of genes with conserved miRNA seed regions.

References

Bartel D: MicroRNAs: target recognition and regulatory functions. Cell 2009, 136(2):215–233.
Khorshid M, Hausser J, Zavolan M, van Nimwegen E: A biophysical miRNA-mRNA interaction model infers canonical and noncanonical targets. Nature Methods 2013
Elati M, Nicolle R, Junier I, Fernandez D, Fekih R, Font J, Kepes F: PreCisIon: PREdiction of CIS-regulatory elements improved by gene's positION. Nucleic Acids Res. 2013 February; 41(3): 1406,1415.
Huang JC, Babak T, Corson TW, Chua G, Khan S, Gallie BL, Hughes TR, Blencowe BJ, Frey BJ, Morris QD: Using expression profiling data to identify human microRNA target. Nature Methods 2007, 4:1045–1049.
Zhijie X, Shiliang S: An Algorithm on Multi-View Adaboost. ICONIP 2010, Sydney, Australia, November 22-25, 2010, Proceedings, Part I, LNCS 6443, pp. 355-362

Title: Integrative biological itemset mining in cancer research

Authors: Naulaerts Stefan[1,2], Vanden Berghe Wim[3], Laukens Kris[1,2]
[1]Department of Mathematics and Computer Science, University of Antwerp, Antwerp, Belgium
[2]Biomedical informatics research center Antwerpen (biomina), University of Antwerp, Antwerp, Belgium
[3]Department of Biomedical Sciences, Lab of Protein Chemistry, Proteomics & Epigenetic Signaling (PPES), University of Antwerp, Antwerp, Belgium

In the last decades, a wealth of data has become available and is ready to be analyzed thanks to careful database design. However, the sheer volume makes it impossible to analyze all this information manually and masks potentially interesting patterns. This problem can partially be overcome by using the advances in machine learning and knowledge discovery as supportive tools for life scientists. Innovative frequent itemset search algorithms are capable of uncovering hidden patterns and can be fine-tuned to focus on capturing biological relevance instead of only the database characteristics. However, this requires that several weaknesses of FIM algorithms will be addressed, such as the dense data problem and the lack of biologically relevant quality measures that can be used to fine-tune the algorithms. Fortunately, many different types of biological information have become available and can be combined to redefine the interestingness criterium to the life sciences environment.

In this poster, we present a frequent itemset mining (FIM) framework, powered by information from public and in-house repositories as an assistant platform for integratomics analyses. As such, we build on existing techniques used in pathway and functional enrichment and combine these with biologically relevant modifications of current state-of-the-art data mining techniques. We hereby tackle several of the traditional shortcomings of FIM algorithms, while validating and applying our multi-level approach to currently ongoing cancer research to identify regulatory systems.

Prediction of genome-wide in vivo transcription factor binding using factor-specific DNase footprinting models

Galip Gürkan Yardımcı[1], Gregory E. Crawford[1,2], Uwe Ohler[1,3]

Keywords: gene regulation, DNase-seq, footprinting, transcription factor binding

The identification of DNase I hypersensitive sites and DNase footprints are well established methods for identification of genomic regulatory regions and DNA-protein interactions, respectively. Using data generated by high throughput DNase-seq assays, we propose models to identify binding locations of transcription factors in different cell lines in a genome-wide manner by modeling each factor's unique DNase footprint. Contrary to most existing approaches, our model aims to represent the footprint shape in detail while trying to account for the contribution of overall DNase hypersensitivity around a binding site to assess the accuracy of the footprints by themselves – a necessary feature to identify specific sites bound under different conditions. We model each transcription factor's footprint using two features: distribution of DNase-seq reads at each base and the DNase-seq coverage. Transcription factor binding predictions are validated rigorously using ChIP-seq assays from the ENCODE consortium. We achieve a mean AUC value of 95% for 20 transcription factors. We find that AUC values tend to depend on quality of motif associated with transcription factor and transcription factor structural family. For each transcription factor, we show that some ChIP-seq peaks do not overlap with a DNase footprint and characterize such peaks according to ChIP-seq signal intensity and co-binding proteins.

[1] Institute for Genome Sciences & Policy, Duke University, Durham, NC

[2] Department of Pediatrics, Duke University, Durham, NC

[3] Department of Biostatistics & Bioinformatics, Duke University, Durham, NC