Abstracts

Regulatory Genomics in Drosophila

Alexander Stark, IMP, Dr. Bohr-Gasse 7, 1030 Vienna, Austria

We are witnessing a tremendous increase in methodology and resources that allow detailed studies of gene regulation. Especially research on the model organism Drosophila melanogster has greatly benefited over the past years, for example from the sequencing of entire genomes for 11 additional Drosophila species and from systematic genome-wide studies, e.g. of transcription factor binding.

In this talk, I will review insights into pre- and post-transcriptional gene regulation enabled by comparative genomics and systematic genome-wide studies and discuss recent approaches and unsolved challenges.

Five-vertebrate ChIP-seq reveals the evolutionary dynamics of transcription factor binding

Benoit Ballester
European Bioinformatics Institute, Hinxton, UK

Transcription factors (TFs) direct gene expression by binding to DNA regulatory regions. To explore the evolution of gene regulation, we used chromatin immunoprecipitation with high-throughput sequencing (ChIP-seq) to determine experimentally the genome-wide occupancy of two TFs, CCAAT/enhancer-binding protein alpha and hepatocyte nuclear factor 4 alpha, in the livers of five vertebrates. Although each TF displays highly conserved DNA binding preferences, most binding is species-specific, and aligned binding events present in all five species are rare. Regions near genes with expression levels that are dependent on a TF are often bound by the TF in multiple species yet show no enhanced DNA sequence constraint. Binding divergence between species can be largely explained by sequence changes to the bound motifs. Among the binding events lost in one lineage, only half are recovered by another binding event within 10 kilobases. Our results reveal large interspecies differences in transcriptional regulation and provide insight into regulatory evolution.

MotifLab: A tools and data integration workbench for motif discovery and regulatory sequence analysis

Kjetil Klepper
Department of Cancer Research and Molecular Medicine, Norwegian University of Science and Technology, Trondheim, Norway

Discovering binding motifs and bindings sites for transcription factors is an important problem in bioinformatics, and many tools have been proposed to search for novel motifs or to scan for potential sites that match established binding motifs. Unfortunately, traditional motif discovery and scanning methods that only rely on sequence data have a tendency to make a lot of false predictions. However, it has been demonstrated that use of additional information, such as gene expression, sequence conservation, location of DNase HS sites and epigenetic marks etc., has the potential to reduce the number of spurious predictions and also discriminate between functional and non-functional binding sites. A lot of data that could prove useful for this purpose is already available at genome-wide scales and more data for different organisms, cell-types and conditions is being published at an increasing rate.

MotifLab is a general software workbench for regulatory sequence analysis designed to make it easy to incorporate different types of data into the motif discovery process. A key application of MotifLab is for constructing positional priors tracks based on various sequence feature annotations. Positional priors can be used to highlight those parts of sequences that are considered more likely to contain functional binding sites, and they can be employed by motif discovery methods to guide the search or be used in a post-processing step to filter unpromising predictions. MotifLab can interface with several popular motif discovery tools (including MEME, Priority, MDscan, Weeder and BioProspector) to predict both individual binding sites and combinations of sites that could potentially function together (cis-regulatory modules).

Identification of regulatory signals using graph modularity analysis

Alexandre P. Francisco[1], Sophie Schbath[2], Arlindo L. Oliveira1 and Ana T. Freitas[1]
[1]*INESC-ID / IST, Tech Univ of Lisbon, Portugal (aplf,aml,atf@inesc-id.pt)*
[2]*INRA, France (sophie.schbath@jouy.inra.fr)*

In the last decade, after the completion of genome sequencing projects for various organisms, the study of gene regulation and gene expression mechanisms has required new computational approaches. Despite the remarkable success of these tools in some areas of application like gene finding, sequence alignment, etc, there are still problems for which no definitive methods have been developed. Notably, the accurate identification of biologically meaningful nucleotide sequences in cis-regulatory regions remains an open problem.

Many algorithms have been proposed to date for the problem of finding biologically significant motifs in promoter regions. They can be classified into two large families: combinatorial methods and probabilistic methods. Probabilistic methods have been used more extensively, since their output is easier to interpret. Combinatorial methods have the potential to identify hard to detect motifs, i.e. both strong and weak signals, but they deluge the user with a large, possibly huge, number of motifs, consisting of hundreds or thousands variations of the motifs of interest.

In this work, we propose a method that can be used to process the output of combinatorial motif finders in order to find groups of motifs that represent variations of the same motif, thus reducing the outputs to manageable sizes. This output processing step is done by building a graph that represents the co-occurrences of motifs. This graph has one node for each motif found, and one edge between two motifs if they have significant occurrence overlap. The identification of motifs is performed by finding communities in the co-occurrences graph using graph clustering techniques. The motifs in each cluster are finally combined into a composed representation, and a position weight matrix (PWM) is generated. Note that our approach is also able to process and identify complex motifs, i.e. motifs that are built of two or more simple motifs, spaced by a number of bases that falls within a specific range.

We show that this innovative approach leads to a method that is as easy to use as a probabilistic motif finder, and as sensitive to low quorum motifs as a combinatorial motif finder. Furthermore, we derive an approach to evaluate the statistical significance of motif clusters, providing motif rankings, and we show how to combine the output of different motif finders, both combinatorial and probabilistic, leading to an innovative integration method which helps to integrate, analyze and compare the output of several motif finders.

Several tests have been preformed against several well known motif finders, using a set of recently published large-scale compendium of transcription factors, derived from diverse high-throughput experiments in several metazoan. The results show that the method is highly competitive with state of the art methods that use much more extensive information.

Modeling regulatory complexes using both TF-DNA and TF-TF interactions

Esko Ukkonen
Department of Computer Science, University of Helsinki, Finland

Transcription factors (TFs) not only bind to DNA but also to each other. The individual interactions between TFs are often weak. However, the complexes formed by both TF-DNA and TF-TF interactions are much more stable than those based on similar TF-DNA interactions alone. Thus, the dimerization mediated by DNA gives a much richer regulatory machinery that has a potentially stronger capability to explain the regulation of gene expression. The talk will describe recent developments in modeling and predicting such regulatory complexes, using data from high-throughput SELEX experiments, and applying this to explain observed binding patterns in DNA.

Joint work with J. Taipale, T. Kivioja, P. Rastas, A. Jolma and J. Toivonen.

The functional importance and detection of regulatory sequence variants

Virginie Bernard
Centre for Molecular Medicine and Therapeutics, Child and Family Research Institute, Department of Medical Genetics - University of British Columbia, Vancouver BC, Canada

The convergence of high-throughput technologies for sequencing individual exomes and genomes and rapid advances in genome annotation are driving a neo-revolution in human genetics. This wave of family-based genetics analysis is revealing causal mutations responsible for striking phenotypes. By mapping the reads to the human genome reference and by searching for variations relative to the reference, a list of small nucleotide variations and structural variations is obtained. Analysis is required to reveal those variations most likely to contribute to a disease phenotype within a family. Existing software such as SIFT and PolyPhen score the severity of changes that arise in protein encoding exons. However, most mutations within a family are situated in the 98% of the genome that controls the developmental and physiological profile of gene activity - the sequences that control when and where a gene will be active.

Functional contributions of cis-regulatory sequence variations to human genetic disease are numerous. With full genome sequencing becoming accessible to medical researchers, the need to identify potential causal mutations in regulatory DNA is becoming imperative. We are implementing a software system to enable genetics researchers to characterize regulatory DNA changes within individual genome sequences. We are combining reference databases of known regulatory elements, experimental archives of protein-DNA interactions and computational predictions within an integrated analysis package. With our software, researchers will have greater capacity to identify variations potentially causal for disease.

The presentation will introduce the challenges and approaches of regulatory sequence variation analysis.

Alternative splicing variability in human populations

Mar Gonzalez-Porta, Micha Sammeth, Miquel Calvo, Roderic Guigo
Center for Genomic Regulation, Barcelona, Catalonia, Spain

We have developed statistical methodology to measure variation in gene expression and splicing ratios within and between populations, and to deconvolute the contribution of each of them to total variability in the abundances of individual transcripts. We have applied this methodology to estimates of transcript abundances obtained from RNA-seq experiments in lymphoblastoid cells from Caucasian and Yoruban individuals. We have found that protein coding genes exhibit reduced gene expression variability in human populations, and an even greater reduction in splicing ratios, with many genes exhibiting constant ratios across individuals. Consistent with this observation, we have found that genes involved in the regulation of splicing show less expression variability than human genes overall. While there is correlation in splicing variability between populations, up to 10% of protein coding genes could exhibit population-specific splicing ratios. We estimate that about 50% of the total variability observed in the abundance of transcript forms can be explained by variability in transcription. A large fraction of the remaining variance can likely result from variability in splicing, although variability in splicing is uncommon without variability in transcription. Genes with high total variability (resulting from variability both in transcription and splicing) are particularly enriched in RNA binding functions. Consistent with this finding (and with the reduced variability of splicing factors), we have also found that long non coding RNAs show higher expression variability than protein coding genes. This suggests that variation in expression of long non coding RNAs may play an important role in establishing the molecular basis of intraspecies phenotypic individuality.

Analyzing and designing small RNA-mediated gene regulation

Pål Sætrom
Dept. of Computer and Information Science and Dept. of Cancer Research and Molecular Medicine, Norwegian University of Science and Technology, Trondheim, Norway

Small non-coding RNAs such as microRNAs (miRNAs) and short interfering RNAs (siRNAs) can regulate many genes by base pairing to sites in mRNAs. For miRNAs, these regulatory targets determine the miRNAs' function, but for siRNAs, most of these regulatory targets represent unwanted side-effects that obfuscate experimental results. Understanding how miRNAs recognize targets is therefore important both for predicting miRNA function and for analyzing siRNA effects. We have developed methods that assess miRNA's overall regulatory effects and have used these to design robust siRNAs.

Transcription factor binding sites, histone modifications, and two promoter classes

Martin Vingron,
Max Planck Institute for Molecular Genetics, Berlin, Germany

This talk will summarize a biophysically motivated method (TRAP) for prediction of transcription factor binding sites and give some applications. Main applications are the identification of tissue specific transcription factors and the prediction of regulatory changes due to SNPs. Further, the talk will describe some indications that the division of promoters into two classes with high and low CpG contents, respectively, is of functional importance and helps in understanding mammalian promoters. In fact, the two classes of promoters display different features when it comes to binding site usage and tissue specific regulation. The dichotomy is further supported by an analysis of histone modifications in the promoters. There we observe that in a log-linear model relating histone modifications to expression level, different sets of modifications are respectively the most informative for prediction in the two classes of promoters. Taken together, we interpret this as indication that different regulatory mechanisms govern transcription in these two classes of promoters.

Use of structural DNA properties for the prediction of regulator binding sites with conditional random fields

Pieter Meysman[1], Thanh Hai Dang[2], Kris Laukens[2], Riet Desmet[1], Yan Wu[1], Kristof Engelen[1] and Kathleen Marchal[1]

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

Molecular recognition of genomic target sites by regulator proteins is a vital process in the transcription regulation of genes in living cells. The types of physical interactions that contribute to the recognition of binding sites by a protein can roughly be divided into those enabling direct read-out and those that allow for indirect read-out [1]. The former comprises base-specific recognition, such as stabilizing hydrogen bonds between regulator amino acids and a set of conserved bases in the genomic DNA sequence, while in the case of the latter variations within the DNA structure will be used as the basis for recognition. It is the direct form of recognition that is the focus of most current endeavors to model regulator binding sites, usually by modeling a conserved set of nucleotides, e.g. a position weight matrix (PWM). However by considering only a single recognition mechanism, these models overlook any information concerning binding site identity that can be derived from the use of indirect read-out by the regulator. It was therefore our goal to create a binding site model based on this second type of recognition which involves interactions between the regulator protein and the molecular structure of the DNA molecule.

Transcriptional cross-regulation as survival mechanism in bacteria

Monsieurs P., Mijnendonckx K., Leys N.,Van Houdt R.
Unit of Microbiology, Belgian Nuclear Research Centre(SCK•CEN), BE-2400, Mol,Belgium.

Abstract

The high number of metal resistance genes in the soil bacterium *Cupriavidus metallidurans* CH34 makes it an interesting model organism to study microbial heavy metal responses. A first step in understanding the molecular mechanisms that underlie heavy metal resistance is to reconstruct the transcriptional regulatory networks. Therefore genomewide expression experiments were performed to investigate the full stress response of *C. metallidurans*CH34 when it was challenged to a variety of heavy metals including zinc, copper, cadmium, and lead. Certain heavy metal response gene clusters showed similar expression profiles when cells were found to be exposed to varying combinations of heavy metals, thus pointing to complex cross-talk at the transcriptional level between the different heavy metal resistance mechanisms. Our results could partially explain this cross-talk by identification and similarity analysis of transcription factor binding sites in the promoter region of metal resistance genes.

This hypothesis is further confirmed by a directed evolution experiment, exposing the bacterium to toxic concentrations of silver, resulting in mutants with an increased resistance towards this metal. Subsequent Illumina sequencing of two of these mutant strains points towards an inactivation of the sensory component of the two-component regulatory system AgrR/S. Remarkably, this regulatory system belongs to a region predicted based on homology to be important for metal resistance(*agrRSCBA*), however without ever being expressed in any transcriptomic study investigating metal resistance. A phylogenetic footprinting approach of the *agrR*promoter region predicts a regulatory motif which resembles the regulatory motifs found in otheridentifiedmetal resistance regions. This way, the AgrR transcription factor would not only activate its own dedicated metal efflux pump AgrCBA, but also other metal resistance regions containing a similar transcription factor binding site in their promoter region, which indeed could be confirmed by gene expression analysis.

A Computational Paradigm for More Specific TFBS Detection

Heike Sichtig [1] and Alberto Riva [1;2]
[1]Department of Molecular Genetics and Microbiology, University of Florida, Gainesville, FL, USA
[2]University of Florida Genetics Institute, University of Florida, Gainesville, FL, USA

Motivation: One of the key challenges of current computational biology is the construction of a model of the regulatory network of a cell. The identification of regulatory patterns in genomic DNA and their relation to specific transcription factors that bind to them is vital to understanding the regulatory infrastructure of a cell.
Our paradigm is based on the combination of two biologically realistic information processing methods: third-generation artificial neural network models (spiking neural networks) are used to represent the complex structure of a binding site, while a genetic algorithm is used to optimize the network parameters during a learning phase. The networks are initially trained using known binding sites and negative examples, and are then used as classifiers to detect new TFBSs in genomic sequences.
The goal of our work is to reduce the number of false positives in the predicted TFBSs, through a more accurate modeling of the information contained in the alignments that constitute the training data.
Results: We present the evaluation of a two-neuron network topology trained to represent TFBSs for four different transcription factors. The networks were trained using real TFBS data from the TRANSFAC, JASPAR and SCPD databases, and appropriately generated negative samples, and were compared against MAPPER, TFBIND and TFSEARCH. Our results show that our paradigm has the potential to attain very high classification accuracy, with a very small number of false positives.

Unveiling combinatorial regulation through the combination of ChIP information and *in silico cis*-regulatory module detection

Hong Sun[1], Tias Guns[2], Siegfried Nijssen[2] and Kathleen Marchal[1,§]

[1]Department of Microbial and Molecular Systems, Katholieke Universiteit Leuven, Address Kasteelpark Arenberg 20, 3001 Leuven, Belgium

[2]Department of Computer Science, Katholieke Universiteit Leuven, Address Celestijnenlaan 200A, 3001 Leuven, Belgium

Introduction

Nowadays with high-throughput chromatin-immunoprecipitation technologies becoming increasingly popular for the genome-wide identification of TF binding sites, *cis*-regulatory module detection (CRM) can be used in combination with ChIP information to computationally predict with which other TFs a ChIP-assayed TF potentially interacts. In contrast to gene centered methods, ChIP information allows reducing largely the regions in which the motif of the assayed TF should be located (typically 500 bp instead of thousands of bp). However, as the binding site of the assayed TF often not coincides with the peak location, searching for CRMs in ChIP-Seq defined regions still boils down to a combinatorial search problem. In addition, as it is not known in advance with which other TF the assayed one interacts, the CRM detection approach needs to be able to search for a CRM that can include any of the known motifs.

Discovery of regulators for co-expressed human genes using large sequence search spaces

Bram Van de Sande, Zeynep Kalender Atak, and Stein Aerts
Laboratory of Computational Biology, Center for Human Genetics, University of Leuven, Belgium.

Introduction

The importance of gene expression profiles provided by microarray and RNA-Seq experiments in disease research, and more specifically in the cancer field, is universally recognized: many studies have shown the added value of this data in the prediction of the prognosis (1) and in our understanding of the pathogenesis of cancer. Knowledge of the causal biochemical mechanisms for these abnormal expression signatures, would even further improve our insights. However, elucidating the perturbed pathways and transcriptional regulation that underlie these aberrant expression profiles remains a challenge.

The gap between gene expression and biochemical pathways is bridged by transcription factors (TF). These TFs bind DNA at specific binding sites that are modeled as Position Weight Matrices (PWM)/motifs. The main challenge can thus be reformulated as the prediction of the motifs that drive a (sub-)set of differentially expressed genes. Classical approaches are based on de novo motif discovery in the proximal promotors of co-expressed genes. However, transcriptional regulation in eukaryotes, and especially in human cells, is complex and also controlled via intronic regions and more distant acting enhancers. Unfortunately, extending the search space to include these more distant regulatory active regions renders classical motif discovery unproductive because of the extended background noise. Therefore, motif discovery is only able to find regulatory signals in search spaces restricted to a region of around 2kb upstream of the TSS. Additionally, most regulatory regions are clustered in Cis-Regulatory Models (CRM) and this kind of. Information is usually not taken into account by motif discovery methods.
For Drosophila melanogaster, a more robust computational analysis technique named CisTargetX (2) is available for the discovery of regulatory motifs that drive a given gene signature. This technique is able to search large genomic spaces for regulatory regions and takes clustering of motifs into account. To overcome the aforementioned problems, we ported the cisTargetX algorithm to handle sets of human genes. This new computational approach will be of great value to the cancer field by enabling researchers to predict the regulatory motifs, and if known, their corresponding TFs, that underlie a given transcriptional aberrant state.

PREFERRED PAIR DISTANCE TEMPLATES REVEAL FUNCTIONAL TRANSCRIPTION FACTOR BINDING SITES

I.V. Kulakovskiy[1,2,*,] A.A. Belostotsky[2], A.S. Kasianov[1], I.A. Eliseeva[4], V.J. Makeev,[2,3]
(1) Engelhardt Institute of Molecular Biology, Russian Academy of Sciences, Vavilov str. 32, Moscow 119991, Russia;
(2) Research Institute for Genetics and Selection of Industrial Microorganisms, 1st Dorozhny proezd 1, Moscow 117545, Russia;
(3) Vavilov Institute of General Genetics, Russian Academy of Sciences, Gubkina str. 3, Moscow 119991, Russia;
(4) Institute for Protein Research, Russian Academy of Sciences, Institutskaya str. 4, Pushchino 142290, Russia.
E-mail: ivan.kulakovskiy@gmail.com

The regulatory code controlling gene expression in higher eukaryotes still remains unclear. It is a complex task to understand how a one-dimensional DNA text of multiple possibly overlapping "words" directs formation of protein complex that controls gene expression in specific tissues in specific conditions. Some insight is given by the well-known concept of "composite elements" consisting of binding sites for different regulatory proteins separated by specific distances [Matys2006]. Despite more than 15 years of study, information about possible intersite distance scale and specificity remains fragmentary.
The quality of identification of transcription factor binding sites (TFBS) both *in vitro* and *in vivo* has increased dramatically with the advent of new technologies like ChIP-Seq. Recently it was observed [Yokoyama2009], [Shelest2010] that there are several preferred distances between some pairs of TFBS (see esp. Fig. 7A in [Yokoyama2009] for pairs of NFY binding sites). Here we illustrate that this phenomenon appears to be much more common at least in the case of *Homo sapiens* TFs involved into regulation of response for hypoxia conditions. We believe this phenomenon can be used to distinguish functional binding sites from false positives, for both experimental identification of TFBS and their prediction *in silico*. We studied distributions of distances between TFBS identified *in silico* for TFs involved in known protein-protein interactions. As a case study we took TFs participating in the regulation of the erithropoetin (EPO) gene expression in hypoxia response in human cells.

Prediction of bacterial small RNA targets

Ivo L. Hofacker
Institut für Theoretische Chemie, Universität Wien, Austria

Bacterial genomes encode a plethora of small RNAs (sRNAs), which are heterogeneous in size, structure, and function. Most sRNAs bind to mRNA targets by means of specific base-pairing interactions, and thereby act as post-transcriptional regulators, modifying translation or stability of the mRNA. The prediction of sRNA targets is therefore a promising avenue to learn about the function of novel sRNAs.

Several approaches for the energy-based prediction of RNA-RNA interactions have been developed over the past years. Ideally, such methods should take into account the competition of intra- and inter-molecular structure formation. Such accessibility based methods are, however, computationally more expensive. I will introduce our improved RNAplex method, which uses an approximate energy model in order to achieve the same time complexity as simple sequence alignment, while still using accessibility. In addition, we will present an RNAplex based web service for the bacterial sRNA target prediction in all currently available bacterial genomes. Finally, we will discuss implications for the number of sRNA targets and the role of sRNA in regulatory networks.

Extended Abstracts