BIOMAGNET ANNUAL MEETING

Brussels, March 21, 2011

Bridging the Gap between Bioinformatics and Modeling

ABSTRACTS

I-DHORE 3.0 DETECTION OF COLLINEARITY IN LARGE SCALE DATASETS.

Sebastian Proost1,2,+ , Jan Fostier3,+, Bart Dhoedt3, Piet Demeester3, Yves Van de Peer1,2, Klaas Vandepoele[1,2] (+contributed equally)

1. Department of Plant Systems Biology, Bioinformatics and Systems Biology Division, VIB, Technologiepark 927, 9052 Ghent, Belgium
2. Department of Plant Biotechnology and Genetics, Ghent University, Technologiepark 927, 9052 Ghent, Belgium
3. Ghent University - IBBT, Department of Information Technology (INTEC), Gaston Crommenlaan 8, Bus 201, 9050 Ghent, Belgium

Abstract

Accurate detection of genomic homology, regions derived from a common ancestor, is imperative to study genome evolution. Collinearity, the conservation of gene content and order is an often used measure for genomic homology. While several tools to detect collinearity exist, few are able to go beyond a pairwise comparison and harvest the information from multiple genomes to detect additional diverged regions. Current tools supporting multiple genomes have memory and run-time issues when scaled up to a dozen or more genomes. As recently several genomes have been released, and many more can be expected in the near future, there is a clear need for a novel tool able to cope with this abundance of date in an efficient way.

Here we present a new version of i-ADHoRe with improved algorithms combined with support for multi-core CPUs and computer clusters. This novel implementation allows analysis of extremely large datasets, including the full Ensemble dataset with 49 genomes, in just a few hours. i-ADHoRe 3.0 clearly shows that the combination of fast heuristics, good implementation and support for modern hardware can be a significant improvement over the current state-of-the-art.

ANNOTATE-IT: A FRAMEWORK AND WEB-INTERFACE FOR THE ARCHIVAL, MANAGEMENT AND INTERPRETATION OF SINGLE NUCLEOTIDE VARIANTS OBTAINED BY NEXT GENERATION SEQUENCING

Alejandro Sifrim[1], Yuching Lai[2], Jeroen Van Houdt[3], Joris Vermeesch[3], Yves Moreau[1], Jan Aerts[1]

1-Katholieke Universiteit Leuven ESAT-SCD Bioinformatics Group, Leuven, Belgium.

2-Leiden University Medical Center, Leiden, Netherlands.

3-Katholieke Universiteit Leuven, Center for Human Genetics, Belgium.

Abstract

Next generation sequencing allows the detection of novel variant sites which are elusive in array-based methods of genotyping using known common variants. As technology develops the throughput increases and the management and interpretation of tens of thousands of variants per sample becomes the new experimental bottleneck. We developed a flexible framework with an easy-to-use interface which allows the archival of variants for every sample and which integrates existing knowledge and prediction software for functional impacts of variants.This allows for the rapid interpretation of NGS results

Our framework is based on a Ruby on Rails system which includes parsing capabilities of most common single nucleotide variant callers. This system is linked to a local Ensembl database to gather exisiting knowledge about the given variants and genes. We use Polyphen2, SIFT and other software to computationally predict the impact of variants on the phenotype. In order to make the task feasible in a reasonable time frame we provide an easy-to-setup parallelized and distributed task management system. The web-interface provides a multi-level complex query system for the archive with the ability to set filters interactively. Visualization of filters and data distributions aid in assessing the impact of filtering in order to choose the most convenient thresholds.

We present the application of Annotate-It in a case-study of full-exome sequencing of 4 patients with Nicolaides-Baraitser syndrome. Doing this we predict the genetic consequences and known information of each variant and predict their possible functional impacts. By doing gene-centered cross-patients queries and taking into account existing knowledge, automatically retrieved by Annotate-It, we can postulate candidate genes for downstream validation.

Due to the sheer magnitude of the datasets produced by NGS experiments, managing and interpreting variant lists becomes a cumbersome and time-consuming task. As more and more data is generated, interesting complex cross-sample questions become more relevant but are hard to achieve. With Annotate-It, we aimed to deliver a tool which handles such data conveniently and easily and which provides a front-end for the biologist or physician to quickly gain new insights from the data. In the future we plan to broaden the scope of Annotate-It further by integrating more knowledge and prediction software.

GENOME PROFILE CORRECTION OF SINGLE-CELL ARRAY CGH USING A STATISTICAL MODEL

Cheng J[1], Konings P[1], Vanneste E[2] ,Vermessch Joris[2] ,Moreau Y[1]

1- Katholieke Universiteit Leuven, ESAT-SCD, Leuven, Belgium

2- Katholieke Universiteit Leuven, Center for Human Genetics, Leuven, Belgium

Abstract

Chromosomal aneuploidy acquired in single cells of human cleavage stage embryos plays an important role in low human fecundity and constitutional chromosomal disorders and tumor progress. Array CGH technologies have been employed to detect chromosomal aneuploidies in single blastomeres on a genome-wide level. However, genome profiles deprived from single cells frequently show a wave pattern, especially at the terminals of a chromosome. Consequently, this genome wave pattern obscures the accurate detection of the real chromosomal aneuploidies and yields high false positive rate (FPR).

The aim of this study was to present a statistical model which efficiently removes the genome wave pattern. The array CGH log2-Ratio was weighted regressed on the corresponding GC percentage. The large weights were assigned to the regions whose GC percentages are large, i.e. GC rich region.

The algorithm was applied to a novel single-cell oligo-array CGH experiment for the copy number variation (CNV) detection. The experiment consisted of 8 single Epstein- Barr virus (EBV) transformed lymphoblastoid cells performed on the Agilent 244K arrays. 12 genomic CNV regions were validated to have CNV by the Affymetrix 250K. Circular Binary Segmentation algorithm was used to compare the CNV detection before and after genome profile correction. The results showed that the genome wave patterns were obviously reduced after the wave correction, in particular for the two terminals of the chromosomes frequently present for many samples. The True Positive Rate (TPR) and the False Positive Rate (FPR) were calculated across 8 EBV cells before and after wave correction. The Mann-Whitney test shows that the TPRs are not significantly different at the 0.05 level whereas the FPR is significantly reduced from median 0.13 to 0.02 at the 0.001 significance level after wave correction. This result indicates that our algorithm has efficiently removed genomic wave patterns without losing the real biological aberrations.

INFORMATION FUSION AND VERTICAL SEARCH BY MULTI-VIEW TEXT MININ

Xinhai Liu1,2, Olivier Gevaert1,2, Léon-Charles Tranchevent1,2, Bart De Moor1,2

1- Department of Electrical Engineering, ESAT-SCD, Katholieke Universiteit Leuven, Kasteelpark Arenberg 10, box 2446, 3001, Leuven, Belgium

2- IBBT-K.U.Leuven Future Health Department, Kasteelpark Arenberg 10, box 2446, 3001, Leuven, Belgium.

Abstract

Biomedical literature contains rich information which could be observed from different point of views and is expected to provide better or exact knowledge about biomedical process. Thus we propose a novel strategy to provide text prior information from multi-view perspectives. The strategy is implemented by text mining on MEDLINE database. Our strategy can be applied to do information fusion by integrating multi-view data or provide certain knowledge from a small vertical perspective.

A Web application of our strategy is developed for gene retrieval. The multiple views can be different controlled vocabularies, weighting schemes, publishing time periods and biomedical subjects. As a series of human related genes are input, based on the multi-view selection, the outputs are the gene-by-term profile, which is illustrated by term cloud and gene-by-gene similariy matrix, which is visualized by color map. In addition, the hierarchical clustering of these queried genes are available as well.

We employ a set of genes which belong to different diseases to test our multi-view gene retrieval system. Firstly, we investigate the dis-similarity of multiple views by calculating the cosine similarity among them and the results demonstrate that multiple views vary with each other, in other words, each of them is able to provide complementary information to certain extend. Secondly, we carry out the hybrid clustering by integrating multi-view text mining data. The clustering results show that integrating multi-view controlled vocabularies and weighting schemes is able to boost clustering performance.

DELAYS IN THE APOPTOTIC SWITCH : LOOK FOR THE HIDDEN SADDLE

Laura Trotta, Rodolphe Sepulchre, Eric Bullinger

University of Liège
Montefiore Institute
4000 Liège (Sart-Tilman)
Belgium

Abstract

Apoptosis, the controlled cell death is a crucial cellular process used by multicellular organisms to remove damaged or potentially harmful cells. Stimulation of a cell with a death-ligand triggers a transition from life to death, leading to the fast activation of specific enzymes called effector caspases. This fast activation is preceded by a variable duration period where only initiator caspases are active. A wide variety of models have been proposed to describe the apoptotic switch and this variable duration delay. Among these models, several use systems of ordinary differential equations to model the interactions between the biochemical components involved in the process. Some of these models are bistable, i.e systems with two stable equilibrium points (life and death). However, some authors argue that bistability is not necessary to reproduce this particular mechanism of switching with delays. There is thus an ongoing debate concerning the necessity or not of bistability for the apoptotic switch.

In this paper, we propose a simple mechanism to create switches with delays. Interestingly, this mechanism is the same regardless of whether the system is bistable or not. The mechanism only relies on the presence of an unstable equilibrium point, a saddle point, with particular properties. We studied three previously published models of apoptosis reproducing the delay period before the fast activation of effector caspases. Two of them are actually bistable and the third one only presents one stable equilibrium point corresponding to death. We show that from a mathematical point of view, these three models share one important characteristic: the three models have a saddle point with specific properties of time-scale separation.

Discussing the mathematical properties of three previously published models of apoptosis, this work shows how complex models achieve the delay period before the fast activation of effector caspases thanks to the presence of a saddle point.

References

Eißing T, Conzelmann H, Gilles E, Allgöwer F, Bullinger E: Bistability analyses of a caspase activation model for receptor-induced apoptosis, *The Journal of Biological Chemistry*, 279(35):36892-36897, 2004.

Albeck J, Burke J, Spencer S, Lauffenburger D, Sorger P: Modeling a Snap-Action, Variable-Delay Switch Controlling Extrinsic Cell Death, *PLoS Biol.,* 6(12):2831-52, 2008

Schliemann M, Eißing T, Scheurich P, Bullinger E : Mathematical modelling of TNF-a induced antiapoptotic signalling pathways in mammalian cells based on dynamic and quantitative experiments, *Proceedings of the 2nd Foundations of Systems Biology in Engineering FOSBE*, 213-218, 2007.

PLAZA 2.0 : A PLATFORM FOR COMPARATIVE GENOMICS IN PLANTS

Michiel Van Bel[1,2,+], Sebastian Proost[1,2,+], Yves Van de Peer[1,2], Klaas Vandepoele[1,2]

1- Department of Plant Systems Biology, VIB, Technologiepark 927, B-9052 Ghent, Belgium.

2 -Department of Molecular Genetics, Ghent University, Technologiepark 927, B-9052 Ghent, Belgium.

+ Contributed equally

Abstract

Comparative sequence analysis has significantly altered our view on the complexity of genome organization and gene function. To explore all this genome information, a centralized infrastructure is required where all data generated by different sequencing initiatives is integrated and combined with advanced methods for data mining.

Here we describe PLAZA 2.0, an update to the PLAZA platform. This resource integrates structural and functional annotation of published plant genomes together with a large set of interactive tools to study gene function and gene, gene family and genome evolution. Containing 23 species for a grand total of more than 840,000 genes, PLAZA 2.0 provides data on most published genomes, over a wide phylogenetic distance. Automatic clustering of gene families was done, and together with subsequent analyses these provide insights in the evolution of orthologs and paralogs. Inter- and intra-colinearity information was computed using I-ADHoRe, providing a valuable view on the ancient history of genomic duplication and speciation events.

PLAZA 2.0 clearly shows how modern large-scale approaches can be used for data mining the majority of published sequenced plant genomes.

THEORETICAL AND EMPIRICAL QUALITY ASSESSMENT OF TRANSCRIPTION FACTOR BINDING MOTIFS

Alejandra Medina-Rivera1,3*, Cei Abreu-Goodger2, Morgane Thomas-Chollier4, Heladia Salgado-Osorio1, Julio Collado-Vides1 and Jacques van Helden1,3

1 . Centro de Ciencias Genómicas, Universidad Nacional Autónoma de México. Av. Universidad s/n. Cuernavaca, Col. Chamilpa, Morelos 62210; Mexico. Email: amedina@lcg.unam.mx, heladia@ccg.unam.mx, collado@ccg.unam.mx

2. EMBL - European Bioinformatics Institute, Wellcome Trust Genome Campus, Hinxton, Cambridge, CB10 1SD, UK. Email: cei@ebi.ac.uk

3. Laboratoire de Bioinformatique des Génomes et des Réseaux (BiGRe). Université Libre de Bruxelles, Campus Plaine, CP 263. Bld du Triomphe. B-1050 Bruxelles, Belgium. Email: Jacques.van.Helden@ulb.ac.be

4. Department of Computational Molecular Biology. Max Planck Institute for Molecular Genetics. Ihnestrasse 73. 14195 Berlin. Germany. Email: thomas-c@molgen.mpg.de

Abstract

We propose a method, implemented in the program matrix-quality, that combines theoretical and empirical score distributions to assess the reliability of position-specific matrices for predicting transcription factor binding sites.

Position-specific scoring matrices are routinely used to predict transcription factor (TF) binding sites in genome sequences. However, their reliability to predict novel binding sites can be far from optimum, due to the use of a small number of training sites or the inappropriate choice of parameters when building the matrix or when scanning sequences with it. Measures of matrix quality such as E-value and information content rely on theoretical models, and may

The program matrix-quality combines theoretical and empirical score distributions to assess the predictive capability of position-specific matrices. The theoretical distribution provides an estimate of the false prediction rate. Empirical distributions indicate the enrichment of binding sites in various collections of sequences: known binding sites (positive control), all upstream regions of a genome, microarray clusters, ChIP-seq peaks. Negative controls are performed by analyzing the same sequence collections with column-permuted matrices.

We applied the method to estimate the predictive capacity of matrices for bacterial, yeast and mouse transcription factors. The evaluation of 60 matrices from RegulonDB revealed some poorly predictive motifs, and allowed us to quantify the improvements obtained by applying multi-genome motif discovery. Interestingly, the results reveal differences between global and specific regulators. It also highlights the enrichment of binding sites in sequence sets obtained from high-throughput ChIP-chip (bacterial and yeast TFs), and ChIP-seq and experiments (mouse TFs).

Users are often restricted by the available databases, which contain motifs of variable quality. In this context, the method presented here has many applications: (i) selecting reliable motifs before scanning sequences; (ii) improving motif collections in transcription factor databases; (iii) evaluating motifs discovered in massive datasets resulting from new sequencing technologies (ChIP-seq, ChIP-chip), to cite a few.

Iterative multi-task sequence labeling for predicting structural properties of proteins

Francis Maes, Julien Becker, Louis Wehenkel

University of Liege - Dept of Electrical Engineering and Computer Science, Institut Montefiore, B28, B-4000, Liege - Belgium.

Abstract

Developing computational tools for predicting protein structural information given their amino acid sequence is of primary importance in protein science. Problems, such as the prediction of secondary structures, of solvent accessibility, or of disordered regions, can be expressed as sequence labeling problems and could be solved independently by existing machine learning based sequence labeling approaches. But, since these problems are closely related, we propose to rather approach them jointly in a multi-task approach. To this end, we introduce a new generic framework for *iterative multi-task sequence labeling*. We apply this - conceptually simple but quite effective - strategy to jointly solve a set of five protein annotation tasks. Our empirical results with two protein datasets show that the proposed strategy significantly outperforms the single-task approaches.

Identification of cis-regulatory elements involved in Zygotic Genome Activation during early *Drosophila melanogaster* embryogenesis.

Elodie DARBO[1], Thomas LECUIT[2], Denis THIEFFRY[3] and Jacques VAN HELDEN[4]

[1] TAGC, UMR628 INSERM, 163 avenue de Luminy, Parc Scientifique de Luminy, 13288, Marseille, Cedex 09, France darbo@tagc.univ-mrs.fr

[2] IBDML, UMR 6216, 63 avenue de Luminy, Parc Scientifique de Luminy, 13288 Marseille Cedex 09, France lecuit@ibdml.univ-mrs.fr

[3] IBENS, INSERM 1024 – CNRS 8187, 46, rue d'Ulm, F-75230 Paris cedex 05, France thieffry@ens.fr

[4] BiGRe, Boulevard du Triomphe, Campus Plaine, Université Libre de Bruxelles, CP263, B-1050 Bruxelles, Belgium jvhelden@ulb.ac.be

In drosophila, the first steps of embryonic development are ensured by maternal mRNAs loaded during oogenesis. No transcription occurs until the first wave of Zygotic Genome Activation (ZGA) at the 8th mitotic cycle (syncytial blastoderm). A second wave occurs at the 14th mitotic cycle, at the time of the cellularisation of the blastoderm.Aiming at deciphering regulatory code governing transcriptional ZGA in Drosophila, we have selected 169 genes activated during ZGA based on transcriptome data [1] and applied motif discovery approaches (Regulatory Sequences Analysis Tools [2], CistargetX [3]) to find over-represented motifs in their non-coding regions. Two known motifs, bound by Zelda [4,5] or Grh [6] and by Trl, respectively, were found by both methods. Three novel motifs were further discovered. We further scanned the non-coding regions of the early induced zygotic genes with the discovered motifs to identify Cis-Regulatory elements Enriched Regions (CRERs).

Finally, following the epigenetic regulatory thread suggested by the discovery of Trl-related motifs, we analyzed recent ChIP-seq datasets for CBP and Trl. Interestingly, the non-coding regions of many early induced zygotic genes were found to be associated with high density of CBP and Trl ChIP-seq reads.

Altogether, these results point towards the implication of several novel factors in ZGA: Trl, CBP, potentially along with other factors binding the discovered motifs with no match in dedicated databases.

To validate these computational predictions, we have selected about thirty CRERs, which will be inserted in reporter constructs to assess their potential regulatory roles during ZGA.

References

[1] F. Pilot, J.M. Philippe, C. Lemmers, J.P. Chauvin, T. Lecuit, Developmental control of nuclear morphogenesis and anchoring by charleston, identified in a functional genomic screen of Drosophila cellularisation. *Development*, 133(4):711-23, 2006.

[2] M. Defrance, R. Janky, O. Sand, J. van Helden, Using RSAT oligo-analysis and dyad-analysis tools to discover regulatory signals in nucleic sequences. *Nat. Protoc.*, 3, 1589-1603, 2008

[3] S. Aerts, X.J. Quan, A. Claeys, M. Naval Sanchez, P. Tate, J. Yan, B. Hassan. Robust target gene discovery through transcriptome perturbations and genome-wide enhancer predictions in Drosophila uncovers a regulatory basis for sensory specification, *PLoS Biology,* 8(7):e1000435, 2010

[4] J.R. ten Bosch, J.A. Benavides, T.W. Cline, The TAGteam DNA motif controls the timing of Drosophila pre-blastoderm transcription. *Development,* 133(10):1967-77, 2006

[5] H.L. Liang, C.Y. Nien, H.Y. Liu, M.M. Metzstein, N. Kirov, C. Rushlow, The zinc-finger protein Zelda is a key activator of the early zygotic genome in Drosophila, *Nature*, 456(7220):400-3 ,2008

[6] M.M. Harrison, M.R. Botchan, T.W. Cline, Grainyhead and Zelda compete for binding to the promoters of the earliest-expressed Drosophila genes. *Dev. Biol.*, 345(2):248-55, 2010

CLINICAL DATA MINER – A GENERIC, MULTI-CENTRIC ELECTRONIC DATA CAPTURE (EDC) FRAMEWORK

Arnaud Installé[1], Thierry Van den Bosch[2], Dirk Timmerman[3], Bart De Moor[4]

(1) Department of Electrical Engineering, ESAT-SCD, Katholieke Universiteit Leuven, Kasteelpark Arenberg 10, box 2446, 3001 Leuven, Belgium
(2) IBBT-K.U.Leuven Future Health Department, Kasteelpark Arenberg 10, box 2446, 3001 Leuven, BelgiumDepartment of Obstetrics and Gynecology, U.Z. Leuven, Herestraat 49, 3000 Leuven,

(3) Department of Electrical Engineering, ESAT-SCD, Katholieke Universiteit Leuven, Kasteelpark Arenberg 10, box 2446, 3001 Leuven, Belgium
(4) IBBT-K.U.Leuven Future Health Department, Kasteelpark Arenberg 10, box 2446, 3001 Leuven, Belgium

Abstract

The aim of this research was to develop a generic, multi-centric Electronic Data Capture (EDC) software framework and web interface that includes features to improve measurement reliability for clinical trials in imaging fields. The software was developed in the context of the studies from the International Endometrial Tumour Analysis (ISUOG) group.

The software framework and web interface were developed using a Test-Driven Development approach, allowing automated quality assurance.

All communication between client and server is encrypted, and the server is protected by a firewall. Daily backups are made and replicated to a remote location.

New studies can be trivially defined using Microsoft Excel Spreadsheets.

The interface (http://cdm.esat.kuleuven.be) supports the use of Visual Analogue Scales (VAS). The "skip pattern" can be used to implement Case Report Forms (CRF) with a hierarchical structure.

Pictures can be shown alongside questions or categories for multiple-choice questions, to assist clinicians in selecting appropriate values for entry fields, which could improve measurement reliability.

The web interface provides auto-completion and interactive validation.

We have created a newly developed multi-centric EDC software framework that can easily be applied to a broad class of clinical studies. The framework is built in a modular way, enabling extensions such as integration in hospital IT-environments.

The web interface is being used for the studies defined by the IETA group.

The possibility to show images next to questions or categories should be of particular interest to studies where imaging modalities have to be evaluated.

USE OF STRUCTURAL DNA PROPERTIES FOR THE PREDICTION OF REGULATOR BINDING SITES WITH CONDITIONAL RANDOM FIELDS

Pieter Meysman[1], Kristof Engelen[1], Kris Laukens[2], Thanh Hai Dang[2] 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.

Abstract

Molecular recognition of genomic target sites by regulator proteins is a vital process in the transcription regulation of genes in living cells. The types of physical interactions that contribute to the recognition of binding sites by a protein can roughly be divided into those enabling direct read-out and those that allow for indirect read-out. The former comprises base-specific recognition, 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.

The structural DNA properties of the binding sites, needed to construct the mod-el, are derived from their nucleotide sequence using a number of higher-order value look-up functions, so-called structural scales, which are based on experimental data (e.g. X-ray crystallography of various DNA molecules). These structural properties were used as input data to train a model representing the common structural features shared by all known binding sites of a specified regulator. This was done using conditional random fields (CRF), a discriminative machine learning method. Two novel extension algorithms were included in the training of the models, namely an optimization method which allows the CRF to work with structural DNA properties, and a correction method which can compensate for any bias in the training set towards nucleotide conservation. Once trained, the models could be used to evaluate the likelihood of regulator binding for any given DNA sequence.

The performance of these models was demonstrated on data sets for 27 different Escherichia coli transcription factors and they show an overall improvement when compared to a standard PWM model and a previously proposed structural property model [1]. Further a set of novel binding site predictions, resulting from use of the trained models in a genome wide screening of E. coli, were validated using a microarray compendium of ca. 1500 arrays, as well as comparison with EcoCyc.

[1] Nikolajewa S, Pudimat R, Hiller M, Platzer M & Backofen R. (2007) BioBayesNet: a web server for feature extraction and Bayesian network modeling of biological sequence data. Nucl Acid Res 35: W688-W693.

CROSS-SPECIES CANDIDATE GENE PRIORITIZATION WITH MERKATOR

Shi Yu[1], Léon-Charles Tranchevent[1], Sonia M. Leach[1], Bart De Moor[1], Yves Moreau[1]

[1] Department of Electrical Engineering ESAT-SCD, Katholieke Universiteit, Leuven, Leuven, Belgium.

Abstract

In biology, there is often the need to prioritize large list of candidate genes to further test only the most promising candidate genes with respect to a biological process of interest. In the recent years, many computational approaches have been developed to tackle this problem efficiently by merging multiple genomic data sources. We have previously described a gene prioritization method based on the use of kernel methods and proved that it outperforms our previous method based on order statistics. In the present poster, we report the extension of the method to support data integration over multiple related species and the development of a web based interface termed MerKator that implements this strategy and proposes candidate gene prioritization for 5 species. Our cross-species approach has been benchmarked and cases studies demonstrate that human prioritizations can benefit from model organism data.

COMPARATIVE MAPPING OF TRANSCRIPTION FACTOR BINDING SITES IN PLANT GENOMES

Elisabeth Wischnitzki[1,2], Yves Van de Peer[1,2], Klaas Vandepoele[1,2]

1. Department of Plant Systems Biology, Bioinformatics and Systems Biology Division, VIB, Technologiepark 927, 9052 Ghent, Belgium
2. Department of Plant Biotechnology and Genetics, Ghent University, Technologiepark 927, 9052 Ghent, Belgium

Abstract

Comparative mapping of transcription factor binding sites results in enrichment for functional binding sites and reduces the number of non-functional and therefore false positive predictions which frequently occur in genome wide motif studies. The growing number of sequenced genomes provides a unique opportunity to study the conservation of transcription factor binding sites in plant species. The detection and evaluation of functional transcription binding sites is essential but unfortunately not trivial. It is especially hindered in plants by several small and large scale duplication events, leading to a very diverse number of orthologous or in many cases inparalogous genes.

Our comparative mapping approach incorporates the individual information of each related set of genes and is applicable to any combination of genomes. To estimate the evolutionary conservation the candidate motifs are evaluated using regulatory regions from orthologous genes. For each gene set and motif the conservation value is calculated as a probability for this event (p-value). The calculation takes the individual composition and size of the set into consideration. Using this approach the p-value can be assigned and the significance of the result is defined. The results are evaluated using published TF target sets and enrichment for functional categories.

Genes with significantly conserved binding sites show higher enrichment of functional categories than a simple genome wide mapping or a naive conservation filtering. Additionally known binding sites lacking functional enrichment using a naive mapping approach can now be linked with the underlying biological process they regulate. The comparison with experimental data shows a clear reduction of false positive instances and increased enrichment for the target sets. Our comparative mapping performed essentially better than simple mapping methods or naïve conservation filtering which sometimes were equal to random.

The results demonstrate that comparative mapping is a powerful approach to identify functional binding sites and to distinguish them from false positives instances. Our approach can also be applied to study gene regulatory network inference and promoter evolution after speciation and/or duplication. This makes it a very effective tool to detect conserved and potentially functional transcription factor binding sites. Especially with the growing number of sequenced genomes the applications of our approach become more distinct and more detailed evolutionary conservation patterns can be studied.

Predicting Receptor-Ligand Pairs through Machine Learning

E. Iacucci 1, F. Ojeda1, B. De Moor1, and Y. Moreau1,†

[1]SCD-ESAT, Department of Electrical Engineering, Katholieke Universiteit Leuven, Leuven, Belgium

† Corresponding author: yves.moreau@esat.kuleuven.be

Abstract

Background Regulation of cellular events is often initiated via extracellular signaling. Extracellular signaling occurs when a circulating ligand interacts with one or more membrane-bound receptors. Identification of receptor-ligand pairs is thus an important and specific form of PPI prediction. Given a set of disparate data sources (expression data, domain content, and phylogenetic profile) we seek to predict new receptor-ligand pairs. We create a combined kernel classifier and assess its performance with respect to the DLRP 'golden standard' as well as the method proposed by Gertz *et al*. [1].

Results Among our findings, we discover that our predictions for the tgfβ family accurately reconstructs over 76% of the supported edges (0.76 recall and 0.67 precision) of the receptor-ligand bipartite graph defined by the DLRP "golden standard". In addition, the combined kernel classifier is able to relatively out-preformed the Gertz *et al*. [1] work by a factor of approximately two as the Gertz *et al*. [1] work reconstruct 44% of the supported edges (0.44 recall and 0.53 precision) of the receptor-ligand bipartite graph defined by the DLRP "golden standard".

Conclusions The prediction of receptor ligand pairings is a difficult and complex task. We have demonstrated that using kernel learning on multiple data sources provides a clear advantage over the existing method in solving this task.

COMPRESSION OF MASS SPECTRAL IMAGING DATA USING DISCRETE WAVELET □ANALYSIS WITH INCORPORATED SPATIAL INFORMATION

Nico Verbeeck[1,2,3], Raf Van de Plas[3,5], Yousef El Aalamat[1,2,3], Bart De Moor[1,2,3] and Etienne Waelkens[3]

1 Katholieke Universiteit Leuven, Dept. of Electrical Engineering (ESAT), SCD-SISTA (BIOI), Kasteelpark Arenberg 10, B-3001 Leuven (Heverlee), Belgium.

2 IBBT-Katholieke Universiteit Leuven Future Health Department, Kasteelpark Arenberg 10, box 2446, 3001 Leuven, Belgium.

3 Katholieke Universiteit Leuven, ProMeta, Interfaculty Centre for Proteomics and Metabolomics, O&N2, Herestraat 49, B-3000 Leuven, Belgium.

4 Katholieke Universiteit Leuven, Dept. of Molecular Cell Biology, O&N, Herestraat 49 - bus 901, B-3000 Leuven, Belgium.

5 Vanderbilt University School of Medicine, Mass Spectrometry Research Center, Nashville, TN.

Abstract

The presented method reduces the size of MSI data sets considerably while still achieving excellent reconstruction of the original mass spectra. This is done by retaining only a limited number of wavelet coefficients that express spatial structure.

Mass Spectral Imaging is a relatively new molecular imaging technology that makes it possible to detect thousands of molecules throughout tissue simultaneously, ranging from low-mass metabolites to high-mass proteins. This technology is of prime interest for the molecular characterization of tissue in biomedical studies.

In recent years, MSI data sets have grown in size to such extent that it becomes more and more infeasible to computationally analyze them in their raw form due to both memory and calculation time constraints.

Previous research at ESAT by Van de Plas et al. has shown solid results using Discrete Wavelet Transform (DWT) on mass spectra to perform feature selection, thus reducing data size, dimensionality and noise. Our newest method further improves on this approach by incorporating one of the key aspects of MSI, spatial information, to better understand what part of the data can truly be considered noise. By using this information, we can selectively remove only those details that do not exhibit a spatial structure.

We demonstrate the performance of this new compression method on a sagittal section of mouse brain and compare the results to the Van de Plas et al. method and to direct analysis of the raw measurements. The presented study focuses on neurodegenerative diseases that show spatially specific behaviour. Examples of such diseases include Parkinson’s disease, where dopamine producing brain nuclei such as the amygdala are affected, and amyotrophic lateral sclerosis, where motor neuron regions in the brain are affected.

By retaining a small number of detail coefficients that express spatial structure we can strongly improve reconstruction of the mass spectra while still achieving considerable compression.

A BENCHMARK ON CANCER CLASSIFICATION USING LS-SVMS AND MICROARRAY DATA

Dusan Popovic[1,2], Anneleen Daemen[1], Bart De Moor[1,2]

1. Katholieke Universiteit Leuven, Dept. of Electrical Engineering (ESAT), SCD-SISTA (BIOI), Kasteelpark Arenberg 10, B-3001 Leuven (Heverlee), Belgium.
2. IBBT-K.U.Leuven Future Health Department, Kasteelpark Arenberg 10, box 2446, 3001, Leuven, Belgium

Abstract

Because methodological choices at multiple steps in the model building process for classification with microarray data can influence performance, a benchmarking study was performed on two cancer data sets (prostate and breast). Different settings for preprocessing, feature selection, the kernel function and kernel parameter optimization scheme were considered for the Least Squares Support Vector Machine classifier. We showed that the cost function of a parameter optimization technique crucially influenced the quality of the derived classifier, while the choice of other components in the model building workflow appeared less important.

LOOKING FOR APPLICATIONS OF MIXTURES OF MARKOV TREES IN BIOINFORMATICS

François Schnitzler[1], Pierre Geurts[1], Louis Wehenkel[1]

[1]. University of Liège, Montefiore Institute & GIGA research

Abstract

The probabilistic graphical model (PGM) framework provides efficient tools to model a probability distribution defined on a large set of variables using models composed of a graph and a set of parameters. The nodes of the graph are often in a one-to-one correspondence with the variables of the problem, and the edges present in the graph are related to the relationships between them, e.g. causal relationships or direct dependency. Parameters are usually grouped in local functions quantifying interactions between variables. As an example, bayesian networks use a directed graph and each local function is a conditional probability distribution of one variable given the variables associated to its parent-nodes in the graph. The product of those conditional distributions is the joint probability distribution over all variables.

Those models are largely used in computational biology, since they provide a visual representation that can be easily understood, can be used to answer a query using the distribution and can be learned automatically from a data set. From a set of gene expression and/or polymorphism data, PGM learning algorithms can infer a regulation network. PGMs have also been trained on databases (with or without expert intervention for specifying the structure) to predict gene position or function, protein structure or interaction loci, or the effect of a molecule on a given pathology. PGMs can also be employed for clustering of biomolecules, or to model time series data.

While PGMs have had already several successful applications in biology, their poor scaling to high dimensional problems (in terms of the number variables) may make them unfit to tackle problems of increasing size. Indeed, both inference and learning are NP-hard on general models, and in practice dealing with thousands of variables is already problematic. The complexity of those two operations is in fact linked by the tree-width of the graphical structure underlying the PGM, i.e. the size of the largest fully connected subgraph of a chordal graph containing that graphical structure, minus one.

It is therefore possible to ensure good algorithmic properties for PGMs by constraining their underlying structures to trees (graph without cycles, tree-width of one), but this constraint limits also the class of problems they are able to model faithfully. Using mixtures of trees, which model a distribution by a weighted sum of tree-structured models, each one defining a distribution on the whole set of variables, leads to improved modeling power while retaining the attractive algorithmic properties, but sacrifices the interpretability of the model.

Our research has so far focused on the development of new methods for learning such mixtures of trees from a data set, for problems with many variables and much fewer samples. Those methods were developed in the perturb and combine framework, where mixtures are constructed by averaging many models built by a randomized learning algorithm, allowing for variance reduction and further improving algorithmic complexity.

Our experiments on synthetic data have shown the interest of these methods, and we now wish to apply them to relevant problems in bioinformatics.

MATCHING AN OSCILLATOR TO ITS PHASE RESPONSE CURVE

Pierre Sacré and Rodolphe Sepulchre[1]

[1] Department of Electrical Engineering and Computer Science, University of Liège, Belgium.

Abstract

Rhythmic phenomena are essential to the dynamic behavior of biological systems. They arise in genetic and metabolic networks as a result of complex interactions between multiple biological processes, which makes their design principles not intuitive. Elucidating those underlying molecular and cellular mechanisms is crucial to advances in systems biology.

Quantitative mathematical models based firmly on experiments provide an essential tool for studying those mechanisms. With recent progresses of biology, the number of key variables involved in a phenomenon increases and the nature of their interactions (feedforward and feedback loops) is better known. In spatially homogeneous conditions, ordinary differential equations describe the time evolution of the system, yet current models suffer from several limitations. Among others, the parameter values are often determined empirically or based on few experimental information.

In mathematical biology, the Phase Response Curve (PRC) has proven a useful input-output tool for the reduction of complex oscillator models. It indicates how the timing of inputs affects the timing (steady-state phase shift) of oscillators. The shape of this curve plays a critical role in entrainment and synchronization properties of the system. Moreover, the PRC is well adapted to description tools developed by biologists. It can be experimentally measured for neurons or circadian rhythms.

We developed a numerical tool to adapt an initial choice of parameters in order to match a particular PRC shape. We built a particular shape-distance pseudometric and computed the gradient of this distance in a point of the parameter space, involving the sensitivity of the PRC. As an application, we study simple biochemical models of circadian oscillators and discuss how sensitivity analysis helps drawing connections between the state-space model of the oscillator and its phase response curve

PREDICTION OF sRNA TARGETS IN *ESCHERICHIA COLI*

Ivan Ishchukov[1], Daniel Ryan[2], Sandra Van Puyvelde[3]

1. ivan.ishchukov@student.kuleuven.be
2 . danielryan9287@gmail.com
3. sandra.vanpuyvelde@biw.kuleuven.be

Abstract

With the upcoming accessibility to high-throughput sequencing and tiling array technologies, an enormous amount of gene expression data becomes publicly available. These data originating from diverse experiments, thereby analyzing particular conditions, can be used for unraveling global regulatory processes. Here we apply a network inference method, LeMoNe, to identify sRNA targets from gene expression data. LeMoNe clusters genes over the different conditions and assigns regulators to modules with co-expressed genes. It uses the expression level of a set of regulators to predict the condition dependent mean expression of the genes in a module. By using different score cut-offs, we were able to analyze sRNA assignments at different stringencies. As input we used a list of 73 sRNAs, some of which with known targets (benchmark) that were assigned to expression modules. Genes in those modules correspond to potential sRNA targets. Those sRNAs which included both those having predicted targets as well as confirmed targets were considered for further analysis. We recovered about 36 % of our benchmark and also predicted 2 novel targets for 2 different sRNAs, which could be validated by literature data.

This *in silico* data is the starting point for further wet lab experiments. The two targets ascribed here will be further analyzed to confirm the direct interaction of the mRNA targets with the sRNAs and to analyze their mode of action. Apart from the assignment of targets to sRNA regulators, this bio-informatics approach provides biologists with information about condition-dependent sRNA regulation, and can help in the design of future experiments.

BOGAS, a wiki style genome annotation portal for eukaryotic genomes.

Lieven sterck[1,2], Kenny Billiau[1,2], Thomas Abeel[1,2], Pierre Rouze[1,2] and Yves Van de Peer[1,2]

1. Department of Plant Systems Biology, VIB, Technologiepark 927, B-9052 Ghent, Belgium.Second address.
2. Department of Molecular Genetics, Ghent University, Technologiepark 927, B-9052 Ghent, Belgium..

Abstract

After the initial gene prediction of a newly sequenced genome is completed, the next step is allowing experts to curate gene models, functional descriptions and to add appropriate comments. BOGAS (BioinformaticsGent Online Genome Annotation Service) is an online annotation portal that allows browsing and on the fly editing of gene descriptions as well as gene structures. It offers through its interface easy access to precomputed information that greatly facilitates the work of a human curator. The portal offers links to related databases (eg. NCBI, JGI, SwissProt, GO, ECnumbers) but also provides access to other (local) online tools such as Blast, and multiple sequence alignment to assist the annotators in the annotation process. Moreover, annotators can make use of fully integrated gene-structure visualisation tools like GenomeView or Artemis to check/modify the proposed gene structures. Upon modification of gene structures all the relevant sequence information, protein and EST alignments, protein domains, … are recalculated and shown in the gene page. All manual annotations are immediately visible for other users.

The system will store all the modification (both functional and structural) from annotators in the database so it can be traced back much like a normal wiki. Because of this wiki-philosophy, the system is ideally suited to allow several people to simultaneously curate initial genome annotations.

Besides the ability to host community annotation efforts, BOGAS has also been equipped with all the necessary features to act as a public genome browser/portal: advanced text-search and Blast functionality as well as a genome browsing interface (AnnoJ).

The BOGAS portal is already hosting several public genomes (*Ectocarpus*,*Ostreococcus*, *Pichia*, poplar and apple) and is currently also being used by several ongoing community annotation projects such as: tomato, *Bathycoccus*, Spidermite, *Eucalyptus* to improve the initial automatic predictions.

BOGAS is available at: http://bioinformatics.psb.ugent.be/webtools/bogas/

Title:

PREDICTING METABOLIC PATHWAYS FROM BACTERIAL OPERONS AND REGULONS

Karoline Faust (1,*), Didier Croes (2), Pierre Dupont (3) and Jacques van Helden (2)

1. Bioinformatics and (Eco-)Systems Biology (BSB). Vrije Universiteit Brussel, Pleinlaan 2. B-1050 Brussels, Belgium. Email: kfaust@vub.ac.be

2. Laboratoire de Bioinformatique des Génomes et des Réseaux (BiGRe). Université Libre de Bruxelles, Campus Plaine, CP 263. Bld du Triomphe. B-1050 Bruxelles, Belgium. Email: Jacques.van.Helden@ulb.ac.be

3. UCL Machine Learning Group, Computing Science and Engineering Department, Universite´ catholique de Louvain, B-1348 Louvain-la-Neuve, Belgium. Pierre.Dupont@uclouvain.be

Abstract:

A number of experimental and bioinformatics analyses results in sets of co-regulated, co-expressed or co-occurring enzyme-coding genes. We used our approach, pathway discovery, to predict metabolic pathways from these enzyme-coding genes, which are assumed to be functionally related. In contrast to pathway matching approaches, pathway discovery, can detect variants or combinations of known metabolic pathways. In addition, it can be applied to organisms whose metabolism is unknown, but for which sets of functionally related, annotated genes are available.

We evaluated the pathway extraction algorithms on 71 MetaCyc pathways and found that a combination of kWalks with a shortest-paths based approach yields the highest accuracy (77%). The pathway extraction tool has been integrated in the Network Analysis Tools (NeAT, http://rsat.ulb.ac.be/neat/), and can be accessed either via a web interface or programmatically. The seed nodes for subgraph extraction can be provided as reactions, (partial) compound names as well as EC numbers or genes. Pathways can be extracted from KEGG, MetaCyc or custom networks.