Highlights from the eleventh ISCB Student Council Symposium 2015

Wilkins, Katie; Hassan, Mehedi; Francescatto, Margherita; Jespersen, Jakob Berg; Gonzalo Parra, R.; Cuypers, Bart; DeBlasio, Dan; Junge, Alexander; Jigisha, Anupama; Rahman, Farzana

Published in:
B M C Bioinformatics

Link to article, DOI:
10.1186/s12859-016-0901-4

Publication date:
2016

Document Version
Publisher's PDF, also known as Version of record

Link back to DTU Orbit

Citation (APA):
Highlights from the 11th ISCB Student Council Symposium 2015

Dublin, Ireland. 10 July 2015

Published: 25 February 2016

A1 Highlights from the eleventh ISCB Student Council Symposium 2015

Katie Wilkins1,2, Mehedi Hassan1, Margherita Francescatto4, Jakob Jespersen2, R. Gonzalez Parra1, Bart Cuypers3,5, Dan DeBlasio1, Alexander Junge1, Anupama Jigisha2, Farzana Rahman3

1Plant Pathology and Plant-Microbe Biology Section, School of Integrative Plant Science, Cornell University, Ithaca, NY, USA; 2Graduate Field of Computational Biology, Cornell University, Ithaca, NY, USA; 3School of Computing & Mathematics, University of South Wales, Cardiff, UK; 4Department of Genome Biology for Neurodegenerative Diseases, German Center for Neurodegenerative Diseases, Helmholtz Association, Germany; 5Department of Surgery, Massachusetts General Hospital, The Broad Institute of MIT and Harvard, Boston, USA; 6Department of Systems Biology, Technical University of Denmark, Kemitorvet, Denmark; 7Protein Physiology Lab, Facultad de Ciencias Exactas y Naturales, Universidad de Buenos Aires, Buenos Aires, Argentina; 8Molecular Parasitology Unit (MPU), Instituto of Tropical Medicine, Nationalstraat 155, 2000 Antwerp, Belgium; 9Advanced Database Research and Modeling (ADReM) research group, University of Antwerp, Middelheimlaan 1, 2000 Antwerp, Belgium; 10University of Arizona, Department of Computer Science, Tucson, AZ 85721, USA; 11Center for non-coding RNA in Technology and Health, Department of Veterinary Clinical and Animal Sciences, University of Copenhagen, Copenhagen, Denmark; 12University College Dublin, Dublin, Ireland

Correspondence: Farzana Rahman (frahman@iscb.org) – School of Computing & Mathematics, University of South Wales, Cardiff, UK

BMC Bioinformatics 2016, Volume 17 Suppl 3

© 2016 The Author(s). This article is distributed under the terms of the Creative Commons Attribution 4.0 International License (http://creativecommons.org/licenses/by/4.0/), which permits unrestricted use, distribution, and reproduction in any medium, provided you give appropriate credit to the original author(s) and the source, provide a link to the Creative Commons license, and indicate if changes were made. The Creative Commons Public Domain Dedication waiver (http://creativecommons.org/publicdomain/zero/1.0/) applies to the data made available in this article, unless otherwise stated.
described a method for dealing with heterogeneity by analyzing multiple models of heterogeneity and calculating model selection statistics to identify the most likely of these models, accounting for noisy data. This method was tested on several real data sets [11]. Although we often think of DNA as a two-dimensional sequence, the development of the high-throughput ChIA-PET method for identifying all physical contacts between distant loci is enabling modeling of DNA as a three-dimensional structure. Przemyslaw Szalaj presented 3D-NOME, a new computational method for modelling the three-dimensional structure of the genome [12]. Based on both ChIA-PET data and known interactions of CTCF and RNAPII, 3D-NOME builds an initial model of the nucleosome in a bottom-up fashion and then uses Monte Carlo simulations to reconstruct each level of structure in a top-down fashion.

Epigenome-wide association studies allow the high-throughput identification of epigenetic markers that contribute to human disease. Charles Edmund Breeze presented eFORGE, a tool for identifying the cell type specificity of differentially methylated positions significantly associated with disease. eFORGE identifies cell types of interest based on enrichment of overlap between significant differentially methylated positions and DNase I hypersensitive sites in each cell type, compared to matched, randomly generated background sets. Jonas Ibn-Salem used Hi-C chromatin conformation to show that paralogous genes share more enhancer elements with one another and is more likely to occur in the same topological association domain than matched, randomly selected gene pairs. This suggests that paralogs share common regulation because and cluster within the three-dimensional chromatin architecture.

As the number of genome sequences available increases, it becomes more challenging for biologists to characterize the data and use it to classify sequenced organisms. Giulia Fiscon presented a new feature extraction method for solving this problem, which she integrated using Monte Carlo simulations to reconstruct each level of structure in a top-down fashion.

Acknowledgements
Because of space constraints we are unable to mention in this publication all the volunteers whose contributions make the Student Council Symposium a reality every year. Our recognition and appreciation goes out to all of them, since without their support the organization of such an event would simply not be possible.

We would like to thank ISCB Executive Director Diane Kovats, ISCB Conferences Director Steven Leard, ISMB Conference Administrator Pat Rodenburg, ISMB programmer Jeremy Henning, and ISCB Administrative Support Suzi Smith for their logistical support and invaluable advice. Furthermore, we thank the ISCB Board of Directors for their continued support of the ISCB Student Council in general and the Student Council Symposium in particular.

We are greatly indebted to ISMB 2015 conference chairs Dr. Alex Bateman and Dr. Janet Kelso for giving us the opportunity to organize the Student Council Symposium 2015 in Dublin.

The Student Council would also like to thank our keynote speakers Dr. Ruth Nussinov and Dr. Des Higgins who generously donated their valuable time by delivering keynote addresses.

The Symposium would not be possible without the financial support of our generous sponsors. We would like to thank BioMed Central, Oxford University Press, Swiss Institute of Bioinformatics, F1000, Bina, and The Genome Analysis Center for their contributions.

We are very grateful to all the volunteer reviewers for their work on ensuring the quality of the scientific program, and to the program and travel fellowship committees for coordinating the reviewing effort.

Finally, we thank all Student Council members that have spent countless hours organizing all aspects of SCS 2015 to ensure its success.

References


O1 Prioritizing a drug’s targets using both gene expression and structural similarity

Griet Laenen1,2, Sander Willems3, Lieven Thorrez2, Yves Moreau1,2
1iMinds Medical IT Department, Leuven, Belgium
2iMinds Medical IT Department, Leuven, Belgium
3Philips Research North America, 345 Scarborough Road, Briarcliff Manor, NY10510, USA

Correspondence: Griet Laenen (griet.laenen@esat.kuleuven.be) – iMinds Medical IT Department, Leuven, Belgium

BMC Bioinformatics 2016, 17(Suppl 3):O1

Background

The pharmaceutical industry is facing unprecedented pressure to increase its productivity. Attrition rates in the later stages of development have risen sharply, with toxicity and lack of efficacy being the main bottlenecks [1]. To address both these safety- and efficacy-related issues, a better understanding of the complex biological response to drug treatment is vital. Although many drugs exert their therapeutic activities through the modulation of multiple targets [2], these targets are often unknown and identification among the thousands of gene products remains difficult. We propose a computational method to support the identification of putative targets of a drug by means of a dual approach combining network diffusion of gene expression with chemical structure similarity.

Methods

The first component of our method prioritizes proteins as potential targets by integrating experimental gene expression data with prior knowledge on protein interactions [3, 4]. More specifically, genes are ranked based on the transcriptional response of functionally related genes by diffusing differential expression signals following treatment over a protein interaction network. In addition, drug-protein interactions can also be predicted from structural information. Building on the similar property principle, the second component of our method prioritizes proteins as drug targets based on the interaction with compounds structurally similar to the drug of interest. To this end compound-compound similarity scores are combined with compound-protein interaction scores. Both this structure-based and expression-based approach produce a genome-wide ranking of potential targets that can eventually be fused to obtain a single ranking.

Results/Conclusion

Our method has been evaluated on a test set of small molecule drugs for which the known targets were derived from ChEMBL [5]. AUC values of up to 90 % were obtained. These results indicate the predictive power of combining gene expression data and structural information for a drug of interest with known protein-protein and protein-compound interaction information respectively, to identify the targets of that drug. As such this dual method can aid in gaining a better knowledge of a candidate drug’s mode of action and its off-target effects and thus be of value in the drug development process.

References


O2 Organism specific protein-RNA recognition: A computational analysis of protein-RNA complex structures from different organisms

Nagarajan Raju1, Sonia Pankaj Chothani1,2,3, Ramakrishnan, Masakazu Sekijima3, Michael Gromiha1
1Department of Biotechnology, Bhuban and Jyoti Mehta School of Biosciences, Indian Institute of Technology Madras, Chennai 600036, Tamilnadu, India; 2Philips Research North America, 345 Scarborough Road, Briarcliff Manor, NY10510, USA; 3Global Scientific Information and Computing Center (GSIC), Tokyo Institute of Technology 2-12-1 Okayama, Meguro-ku, Tokyo 152-8550, Japan

Correspondence: M. Michael Gromiha (gromiha@iitm.ac.in) – Department of Biotechnology, Bhuban and Jyoti Mehta School of Biosciences, Indian Institute of Technology Madras, Chennai 600036, Tamilnadu, India

BMC Bioinformatics 2016, 17(Suppl 3):O2
Motivation
Protein-RNA interactions play essential roles in many cellular processes. It is unclear whether same RNA binding proteins from different organisms show unique patterns or variations to recognize RNA. To address this issue, we have constructed 18 sets of same protein-RNA complexes belonging to different organisms and analyzed the interactions and interacting patterns using various sequence and structure based features [1].

Results
We have investigated the recognizing elements by grouping the protein chains into five organisms such as E. coli, H. sapiens, S. cerevisiae, archaea and thermophiles [2]. We observed that positively charged residues are highly preferred in E. coli whereas aromatic residues and polar residues show preference in S. cerevisiae and thermophiles, respectively. In case of RNA, adenine and uracil are highly preferred in H. sapiens and S. cerevisiae, respectively. The neighboring residues around the binding sites are unique in different organisms (eg. Cys-His in H. sapiens, Ala-leu in E. coli, Ser-Arg in S. cerevisiae, Gly-Arg in archaea and Asn-Lys in thermophiles). Further, molecular dynamics simulations of aspartyl tRNA synthetase complexes from E. coli, T. thermophilus and S. cerevisiae revealed the similarities and differences in structurally equivalent binding site residues to understand the recognition mechanism.

Conclusion
Sequence and structural analysis along with MD simulation showed the variations in the interactions between protein and RNA to understand the recognition mechanism of protein-RNA complexes.

References

Fig. 1 (abstract O3). MCMC fit of two-state diffusion model with measurement noise to an LFA-1 trajectory, from [5]. Colour denotes inferred diffusion state, with green slow diffusion and blue fast diffusion. Colorbar length 100 nm
In this work we describe both the model construction and simulation steps of our algorithm. We do also highlight main advantages of our approach compared to existing methods for genome architecture modeling. We hope that further refinement of 3D-NOME and application to additional ChIA-PET and other types of 3D genome mapping data will help to advance our understanding of genome structural organization and functioning.

References

O5 A novel feature selection method to extract multiple adjacent solutions for viral genomic sequences classification
Giulia Fiscon1,2, Emanuel Weitschek1,2, Massimo Ciccozzi3, Paola Bertolazzi2, Giovanni Felici2
1Department of Computer, Control and Management Engineering, Sapienza University, 00185-Rome, Italy; 2Institute for Systems Analysis and Computer Science, National Research Council, 00185-Rome, Italy; 3Department of Engineering, International University UNINETTUNO, 00186-Rome, Italy
Correspondence: Giulia Fiscon (fiscon@dis.uniroma1.it) – Institute for Systems Analysis and Computer Science, National Research Council, 00185-Rome, Italy
BMC Bioinformatics 2016, 17(Suppl 3):O5

Background
Leveraging improvements of next generation technologies, genome sequencing of several samples in different conditions led to an exponential growth of biological sequences. However, these collections are not easily treatable by biologists to obtain a thorough data characterization and require a high cost-time investment. Therefore, computing strategies and specifically automatic knowledge extraction methods that optimize the analysis focusing on what data are meaningful and should be sequenced are essential [1].

Methods
Here, we present a new feature-selection algorithm based on mixed integer programming methods [2] able to extract multiple and adjacent solutions for supervised learning problems applied to biological data. We focus on those problems where the relative position of a feature (i.e., nucleotide locus) is relevant. In particular, we aim to find sets of distinctive features, which are as close as possible to each other and which appear with the same required characteristics. Our algorithm adopts a fast and effective method to evaluate the quality of the extracted sets of features and it has been successfully integrated in a rule-based classification framework [3].

Results
Our algorithm has been applied to three viral datasets (i.e., Rhino-, Influenza-, Polyomaviruses [4-6]) and enables to extract all the alternative solutions of virus specimen to species assignments, by identifying portions of sequence that are discriminant, compact, and as shorter as possible.

To conclude, we succeeded in extracting a wide set of equivalent classification rules, focusing on short regions of sequences with high reliability and low computational time, in order to provide the biologists with short and highly informative genome parts to be sequenced, as well as a powerful instrument both scientifically and diagnostically, e.g., for automatic virus detection.

References
Jean-Claude Dujardin and Kris Laukens contributed equally to this work.

Background

The protozoan parasite *Leishmania donovani* is the cause of visceral leishmaniasis in the Indian subcontinent and poses a threat to public health due to increasing drug resistance. Only little is known about its peculiar molecular biology and the ‘omics integration efforts conducted so far are very limited. Here we present an integratory database that contains all genomics, transcriptomics, proteomics and metabolomics experiments that are currently publicly available for *Leishmania donovani*. In addition, the user interface contains new analysis tools that uses powerful pattern mining strategies like frequent itemset mining algorithms to detect which proteins and metabolites frequently exhibit the same behaviour under different conditions. Another tool converts several –omics layers in new datasets.

Methods

We developed a user friendly tool to crosslink all existing *L. donovani* –omics experiments. Genomics, transcriptomics, proteomics, metabolomics and phenotypic data were collected and added to a MySQL database compendium, further complemented with publicly available data. Relations between different ‘omics layers were explicitly defined and provided with a level of confidence. Python scripts were developed to preprocess, analyse and import the data. To allow comparability between different experiments the principles of the COLOMBOS bacterial expression compendium were adapted [1].

Next to this vast data resource, a set of integrative data-analysis tools was developed based on data mining strategies. For example: One tool uses frequent itemset mining algorithms to detect which proteins and metabolites frequently exhibit the same behaviour under different conditions. Another tool converts several –omics layers to a network format that can be opened in Cytoscape [2] thus be the basis for network analysis. Django and Twitter Bootstrap frameworks were used to create a web portal to make the tools accessible to any Leishmania researcher.

Results

Excellent public gene, protein and metabolite annotation databases are already available for *Leishmania* (e.g. TriTrypDB [3] and GeneDB [4]). However, the added value of our tool is that it links these annotation data to ‘omics experiments that are either provided by the user, or publicly available. New experiments can quickly be preprocessed, analysed and integrated in the database via its Python back end. Using the compendium and its tools, we characterized the development and drug-resistance of *Leishmania donovani* in a system biology context. The genomes of more than 200 strains were examined for associations with phenotypical features and a subset was linked to transcriptomics, proteomics and metabolomics results. The compendium and its scripts were designed to be generic and can therefore be used for other organisms with only minor adaptations to the original setup.

References


O7 Unravelling signal coordination from large scale phosphorylation kinetic data

Westa Domanova1,2, James R. Krycer1,3, Fatemeh Vafaee1,3, David E. James1,3, Zdenka Kuncic1,2,3,4.

1 Charles Perkins Centre, The University of Sydney, NSW 2006, Australia; 2School of Physics, The University of Sydney, Sydney, NSW 2006, Australia; 3School of Molecular Bioscience, The University of Sydney, NSW 2006, Australia; 4School of Mathematics and Statistics, The University of Sydney, NSW 2006, Australia; 5NIH/NIEHS, NC 27709, USA; 6Department of Proteomics and Signal Transduction, Max Planck Institute for Biochemistry, Martinsried, 82152, Germany

Correspondence: David E. James – Sydney Medical School, The University of Sydney, NSW 2006, Australia; Zdenka Kuncic (zdenka.kuncic@sydney.edu.au) – School of Physics, The University of Sydney, Sydney, NSW 2006, Australia

BMC Bioinformatics 2016, Volume 17 Suppl 3

Page 208 of 209
accessible repository of experimentally-predicted kinase substrate relationships. Next, based on the substrates reported for each kinase in this database, we identify how the kinase activities change over time in temporal datasets.

**Results**

Applying this to an insulin-stimulated phosphorylation screen we were able to distinguish between the substrates of AKT and RPS6KB1, two kinases with the same consensus motif, and identified IRS-1-S270 as a novel putative AKT site. We subsequently used our ssKSR-LIVE algorithm to predict novel substrates for the kinases driving insulin signaling, shedding light on their role in driving insulin-stimulated biological processes. This algorithm can be applied to other high-throughput screens of signal transduction, and thus can be used to improve our understanding of complex diseases caused by dysregulated signalling, including cancer and type 2 diabetes.