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

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 Wilkins 1,2, Mehedi Hassan 3, Margherita Francescatto 4, Jakob Jespersen 5,6, R. Gonzalez Parra 1, Bart Cuypers 5,6, Dan DeBlasio 7, Alexander Junge 1, Anuparna Jigisha 2, Farzana Rahman 3

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 (DZNE) within the 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), Institute of Tropical Medicine, Nationalestraat 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

Katie Wilkins and Farzana Rahman contributed equally to this work.

Abstract

This report summarizes the scientific content and activities of the annual symposium organized by the Student Council of the International Society for Computational Biology (ISCB), held in conjunction with the Intelligent Systems for Molecular Biology (ISMB) / European Conference on Computational Biology (ECCB) conference in Dublin, Ireland on July 10, 2015.

About the Student Council and the symposium

The Student Council (SC), part of the International Society for Computational Biology (ISCB), aims at nurturing and assisting the next generation of computational biologists. Our membership and leadership are composed of volunteer students and post-docs in computational biology and related fields. The main goal of our organization is to offer networking and soft skill development opportunities to our members.

The Student Council Symposium (SCS) takes place every year, directly preceding the ISMB/ECCB conferences. SCS 2015 marked the eleventh consecutive edition of the event [1-7].

Meeting format

The Student Council Symposium is a one-day event. Repeating a successful event from previous years, SCS 2015 opened with a scientific speed dating session. During this session our delegates find a partner to introduce themselves to and then learn about each other’s scientific backgrounds and interests. After ten minutes everyone switches partners, and this goes on until the allotted time runs out. The traditional scientific component of the meeting consisted of two keynote presentations by senior researchers, twelve student presentations, and a poster session.

At SCS 2015, Dr. Ruth Nussinov (National Cancer Institute, USA and Tel Aviv University, Israel) and Dr. Des Higgins (Conway Institute, University College Dublin, Ireland) generously agreed to deliver the keynote addresses. The symposium also included two short presentations on open science. The first, about publishing in the digital era, was given by Dr. Michael Markie, a faculty member at our institutional partner F1000 (UK). The second, about data sharing, was given by Dr. Robert Davey, a group leader with our institutional partner The Genome Analysis Center (UK).

Students submitted 100 abstracts to SCS 2015, which were peer-reviewed by 25 independent reviewers. Approximately 75 abstracts were accepted for poster presentations and 12 of these were also accepted for oral presentation. Extended abstracts of oral presentations are included in this report. All abstracts are available online in the SCS 2015 booklet (http://scs2015.iscbsc.org/scs2015-booklet).

Keynotes

The day opened with Dr. Ruth Nussinov’s keynote, in which she explained that Ras GTPase proteins involved in signal transduction include oncogenes such as KRas4B. While little is currently known about the mechanism by which the different mutations of KRas4B lead to cancer, Dr. Nussinov revealed that different mutations are differentially associated with different types of cancer. She believes that “structural biology, computations and experiment, are uniquely able to tackle” this question.

In the afternoon, Dr. Des Higgins began his keynote presentation with a history of multiple alignment algorithms. He then presented his newest alignment program, Clustal Omega, designed to align large numbers of sequences quickly and accurately.

Student presentations

The student presentations were begun by Griet Laenen, who shared a new method for identifying drug targets [8]. Each putative target is ranked based on the transcriptional response of functionally related genes and based on the structural similarity of the drug to proteins know to interact with the putative target. On a ChEMBL-derived test set, AUC values of up to 90 % were achieved.

Nagarajan Raju reported the identification of different modes of protein-RNA interaction in different organisms, based on structural analysis and molecular dynamics simulation [9]. The long reads available from third generation sequencing methods are a valuable resource, but can complicate mapping because of their higher error rate. Hybrid methods use more accurate short reads to correct these errors but current methods for doing so are either too slow or less accurate on larger genomes. Giles Miclotte presented the Jabba hybrid error correction method which uses corrected de Bruijn graphs to achieve comparable performance on small genomes while still performing well for larger genomes [10].

Drawing robust biological conclusions from stochastic single particle tracking trajectories is particularly difficult when particle movement is heterogeneous, such as in the plasma membrane. Paddy Slator
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. Przemysław Szalai 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 virus specimens to viral species [13]. Charles Edmund Breeze presented eFORGE, a tool for identifying the virus specimens to viral species [13].

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 virus specimens to viral species [13].

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 virus specimens to viral species [13].

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 virus specimens to viral species [13].

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 virus specimens to viral species [13].

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 virus specimens to viral species [13].

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 virus specimens to viral species [13].

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 virus specimens to viral species [13].

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 virus specimens to viral species [13].

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 virus specimens to viral species [13].

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 virus specimens to viral species [13].

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 virus specimens to viral species [13].

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 virus specimens to viral species [13].

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 virus specimens to viral species [13].

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 virus specimens to viral species [13].

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 virus specimens to viral species [13].

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 virus specimens to viral species [13].

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 virus specimens to viral species [13].

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 virus specimens to viral species [13].

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 virus specimens to viral species [13].

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 virus specimens to viral species [13].

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 virus specimens to viral species [13].

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 virus specimens to viral species [13].

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 virus specimens to viral species [13].

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 virus specimens to viral species [13].

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 virus specimens to viral species [13].

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 virus specimens to viral species [13].

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 virus specimens to viral species [13].

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 virus specimens to viral species [13].

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 virus specimens to viral species [13].

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 virus specimens to viral species [13].

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 virus specimens to viral species [13].

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 virus specimens to viral species [13].

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 virus specimens to viral species [13].

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 virus specimens to viral species [13].

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 virus specimens to viral species [13].

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 virus specimens to viral species [13].

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 virus specimens to viral species [13].

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 virus specimens to viral species [13].

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 virus specimens to viral species [13].

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 virus specimens to viral species [13].

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 virus specimens to viral species [13].

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 virus specimens to viral species [13].

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 virus specimens to viral species [13].

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 virus specimens to viral species [13].

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 virus s...


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

Griet Laenen1,2, Sander Willems3, Lieven Thorez4, Yves Moreau1,2
1KU Leuven, Dept. of Electrical Engineering (ESAT), STADIUS Center for Dynamical Systems, Signal Processing and Data Analytics, Leuven, Belgium; 2iMinds Medical IT Dept, Leuven, Belgium; 3Scientific Institute of Public Health (IPH), Platform of Biotechnology and Molecular Biology (PBB), Brussels, Belgium; 4KU Leuven, Dept. of Development and Regeneration @ Kulak, Kortrijk, Belgium

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 Chothani2,3,4,5,6, Ramakrishnan1, Masakazu Sekijima7, Michael Gromiha1
1Department of Biotechnology, Bhupat and Jyoti Metha 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 Ookayama, Meguro-ku, Tokyo 152-8550, Japan; 4Global Scientific Information and Computing Center, Tokyo Institute of Technology 2-12-1 Ookayama, Meguro-ku, Tokyo 152-8550, Japan; 5Department of Biotechnology, Bhupat and Jyoti Metha School of Biosciences, Indian Institute of Technology Madras, Chennai 600036, Tamilnadu, India

Correspondence: M. Michael Gromiha (gromiha@bitm.ac.in) – Department of Biotechnology, Bhupat and Jyoti Metha 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 the same RNA binding proteins from different organisms show unique patterns or variations to recognize RNA. To address this issue, we have conducted 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 (E.g. Cys-His in H. sapiens, Ala-Leu in E. coli, Ser-Arg in S. cerevisiae, Gly-Arg in archaea and Asp-Lys in thermophiles). Further, molecular dynamics simulations of aspartyl tRNA synthetase complexes from E. coli, T. thermophila 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

---

**O3**

**Detection of Heterogeneity in Single Particle Tracking Trajectories**

Paddy J Slator1,2*, Nigel J Burroughs1

1Systems Biology Centre, University of Warwick, Senate House, Coventry, UK; 2Systems Biology Doctoral Training Centre, University of Warwick, Senate House, Coventry, UK

Correspondence: Paddy J Slator (p.j.slator@warwick.ac.uk) – Systems Biology Doctoral Training Centre, University of Warwick, Senate House, Coventry, UK

**BMC Bioinformatics** 2016, 17(Suppl 3):O3

Background
Single particle tracking trajectories are fundamentally stochastic, which makes the extraction of robust biological conclusions difficult. This is especially the case when trying to detect heterogeneous movement of molecules in the plasma membrane. This heterogeneity could be due to a number of biophysical processes such as: receptor clustering [1], traversing lipid microdomains [2, 3] or cytoskeletal barriers [4].

Results
Working in a Bayesian framework, we developed multiple models for heterogeneity, such as confinement in a harmonic potential well, switching between diffusion coefficients and diffusion in a fenced environment (or ‘hop’ diffusion). We implement these models using a Markov chain Monte Carlo (MCMC) methodology, developing algorithms that infer model parameters and hidden states from single trajectories. We also calculate model selection statistics, to determine the most likely model given the trajectory. Our methodology also accounts for measurement noise. For LFA-1 receptors diffusing on T cells we previously showed that 12-26 % of trajectories display clear switching between diffusive states, depending on treatment (example trajectory in Fig. 1) [5].

Analysis of the motion of GM1 lipids bound to the cholera toxin B subunit in model membranes confirmed transient trapping in harmonic potential wells. We developed an algorithm which detects hopping diffusion, and validated on simulated data (Fig. 2). We have also demonstrated that allowing for measurement noise is essential, as otherwise false detection of heterogeneity may be observed.

Conclusions
We have used Bayesian methodology to analyze single particle tracking trajectories. Rather than methods which rely on generic properties of Brownian motions, our approach allows us to test which biophysical model best fits a trajectory. With the continuing improvement in spatial and temporal resolution of trajectories, these methods will be important for biological interpretation of single particle tracking experiments.

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](image-url)
In this work we describe both the model construction and simulation approach compared to existing methods for genome architecture modeling. We demonstrate the effectiveness of 3D-NOME in building 3D genome models at the kilobase resolution using ChIA-PET data of human B-lymphocytes (GM12878 cell line) [1].

We also consider CTCF-motif orientations in order to obtain more reliable structures. The specificity of the ChIA-PET data allows us to model the shape of individual chromatin loops and their mutual interactions within topological domains boundaries. We demonstrate the effectiveness of 3D-NOME in building 3D genome models at the kilobase resolution using ChIA-PET data of human B-lymphocytes (GM12878 cell line) [1].

Conclusion

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 Fison1,2, Emanuel Weitschek2,3, Massimo Ciccozzi4, Paola Bertolazzi2, Giovanni Felici2

1Department of Computer, Control and Management Engineering, Sapienza University, 00185-Rome, Italy; 2Istitute for Systems Analysis and Computer Science, National Research Council, 00185-Rome, Italy; 3Department of Engineering, International University UNINETTUNO, 00186-Rome, Italy; 4Istituto Superiore di Sanità, 00161-Rome, Italy

Correspondence: Giulia Fison (fison@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
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 data-compendium, further complemented with publicly available omics experiments that are currently publicly available for Leishmania donovani. In addition, the user interface contains new analysis tools that use powerful pattern mining strategies like frequent itemset mining to allow the linking of results from different 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


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.