ECCB 14
STRASBOURG · FRANCE
7-10 SEPTEMBER 2014

Program of the Wednesday, September 10

Wednesday, September 10

8:00	Conference registration opens	Main Floor
9:00	**Keynote 6: Doron LANCET**. ***Rational confederation of genes and diseases***.	Auditorium ERASME
	Session Wed1: Databases and Ontologies	
9 :50	*PP34 - The impact of incomplete knowledge on the evaluation of protein function prediction: a structured-output learning perspective*. Yuxiang Jiang	
10:15	*PPE35 - Integration of molecular network data reconstructs Gene Ontology*. Vladimir Gligorijevic	
10 :40	Coffee Break	Main Floor
	Session Wed2: Bioinformatics of Health and Disease (2) and Bio-Imaging	Auditorium ERASME
11:00	*PP36 - Transcriptome-guided amyloid imaging genetic analysis via a novel structured sparse learning algorithm.* Jingwen Yan	
11:25	*PP37 - Large-scale automated identification of mouse brain cells in confocal light sheet microscopy images*. Paolo Frasconi	
11:50	**Highlight Talk:** *HP09 - Novel Developments in computational clinical breath analysis and biomarker detection*. Anne Christine Hauschild	
12:15	Moving to Dining Room	
12:30	LUNCH	Dining Room Contades
13:30	**Industrial and Demo Track**	See corresponding pages
	Session Wed3: Text Mining for Computational Biology	Auditorium Erasme
14:35	*PP38 - Extracting patterns of database and software usage from the bioinformatics literature*. Geraint Duck	
15:00	**Highlight Talk:** *HP10 - Text mining technologies for database curation*. Fabio Rinaldi	
	Session Wed4: RNA prediction	Auditorium Erasme
15:25	*PP39 - CRISPRstrand: Predicting repeat orientations to determine the crRNA-encoding strand at CRISPR loci*. Omer S. Alkhnbashi	
15:50	*PP40 - Towards a piRNA prediction using multiple kernel fusion and support vector machine.* Fariza Tahi.	
16:15	**Keynote 7: Eric WESTHOF**. ***The Detection of Architectural Modules in RNA sequences and the RNA-Puzzles Modeling Contest.***	
17:05	Prize Ceremony and Concluding remarks	
17:15	End of Conference – Farewell Coffee	Main Floor

Wed1 (Area H): Biological Ontologies

Chairs: To be announced

PP34 - The impact of incomplete knowledge on the evaluation of protein function prediction: a structured-output learning perspective

Yuxiang Jiang[1], Wyatt Clark[1], Iddo Friedberg[2,3] and Predrag Radivojac[1]

[1]Department of Computer Science and Informatics, Indiana University, Bloomington, Indiana, USA. [2]Department of Microbiology, Miami University, Oxford, Ohio, USA. [3]Department of Computer Science and Software Engineering, Miami University, Oxford, Ohio, USA.

ABSTRACT

Motivation: The automated functional annotation of biological macro-molecules is a problem of computational assignment of biological concepts or ontological terms to genes and gene products. A number of methods have been developed to computationally annotate genes using standardized nomenclature such as Gene Ontology (GO). However, questions remain about the possibility for development of accurate methods that can integrate disparate molecular data as well as about an unbiased evaluation of these methods. One important concern is that experimental annotations of proteins are incomplete. This raises questions as to whether and to what degree currently available data can be reliably used to train computational models and estimate their performance accuracy.

Results: We study the effect of incomplete experimental annotations on the reliability of performance evaluation in protein function prediction. Using the structured-output learning framework, we provide theoretical analyses and carry out simulations to characterize the effect of growing experimental annotations on the correctness and stability of performance estimates corresponding to different types of methods. We then analyze real biological data by simulating the prediction, evaluation, and subsequent re-evaluation (after additional experimental annotations become available) of GO term predictions. Our results agree with previous observations that incomplete and accumulating experimental annotations have the potential to significantly impact accuracy assessments. We find that their influence reflects a complex interplay between the prediction algorithm, performance metric, and underlying ontology. However, using the available experimental data and under realistic assumptions, our results also suggest that current large-scale evaluations are meaningful and almost surprisingly reliable.

Contact: predrag@indiana.edu

Supplementary information: Supplementary data are available at Bioinformatics online.

PP35 - Integration of molecular network data reconstructs Gene Ontology

Vladimir Gligorijevic, Vuk Janjic and Natasa Przulj

Department of Computing, Imperial College London, SW7 2AZ, UK.

ABSTRACT

Motivation: Recently, a shift was made from using Gene Ontology (GO) to evaluate molecular network data to using these data to construct and evaluate GO: Dutkowski et al. [2013] provide the first evidence that a large part of GO can be reconstructed solely from topologies of molecular networks. Motivated by this work, we develop

a novel data integration framework that integrates multiple types of molecular network data to reconstruct and update GO. We ask how much of GO can be recovered by integrating various molecular interaction data.

Results: We introduce a computational framework for integration of various biological networks using Penalized Non-negative Matrix Tri-Factorization (PNMTF). It takes all network data in a matrix form and performs simultaneous clustering of genes and GO terms, inducing new relations between genes and GO terms (annotations) and between GO terms themselves. To improve the accuracy of our predicted relations, we extend the integration methodology to include additional topological information represented as the similarity in wiring around non-interacting genes. Surprisingly, by integrating topologies of bakers yeasts protein-protein interaction, genetic interaction and co-expression networks, our method reports as related 96% of GO terms that are directly

related in GO. The inclusion of the wiring similarity of non-interacting genes contributes 6% to this large GO-term association capture. Furthermore, we use our method to infer new relationships between GO terms solely from the topologies of these networks and validate 44% of our predictions in the literature. In addition, our integration

method reproduces 48% of cellular component, 41% of molecular function and 41% of biological process GO terms, outperforming the previous method in the former two domains of GO. Finally, we predict new GO annotations of yeast genes and validate our predictions through genetic interactions profiling.

Supplementary information: Supplementary Tables of new GO term associations and predicted gene annotations are available at: http://bio-nets.doc.ic.ac.uk/GO-Reconstruction/ .

Contact: natasha@imperial.ac.uk

Wed2 (Area G): Bioinformatics of Health and Disease (2) and Bio-imaging

Chair: Lodewyk Wessels

PP36 - Transcriptome-guided amyloid imaging genetic analysis via a novel structured sparse learning algorithm

Jingwen Yan[1,2], Lei Du[2], Sungeun Kim[2], Shannon Risacher[2], Heng Huang[3], Jason Moore[4], Andrew Saykin[2] and Li Shen[2], and the Alzheimer's Disease Neuroimaging Initiative[§]

[1]BioHealth, Indiana University School of Informatics & Computing, Indianapolis, IN, 46202, USA. [2]Radiology & Imaging Sciences, Indiana University Sch. of Medicine, Indianapolis, IN, 46202, USA. [3]Computer Science & Engineering, The University of Texas at Arlington, TX, 76019, USA. [4]Genetics, Community & Family Medicine, Dartmouth Medical School, Lebanon, NH, 03756, USA. [§]A complete listing of ADNI investigators can be found at: http://adni.loni.usc.edu/wp-content/uploads/how to apply/ADNI Acknowledgement List.pdf.

ABSTRACT

Motivation: Imaging genetics is an emerging field that studies the influence of genetic variation on brain structure and function. The major task is to examine the association between genetic markers such as single nucleotide polymorphisms (SNPs) and quantitative traits (QTs) extracted from neuroimaging data. The complexity of these data sets have presented critical bioinformatics challenges that require new enabling tools. Sparse canonical correlation analysis (SCCA) is a bi-multivariate technique used in imaging genetics to identify complex multi-SNP-multi-QT associations. However, most of the existing SCCA algorithms are designed using the soft thresholding method, which assumes that the input features are independent from one another. This assumption clearly does not hold for the imaging genetic data. In this paper, we propose a new knowledge-guided SCCA algorithm (KG-SCCA) to overcome this limitation as well as improve learning results by incorporating valuable prior knowledge.

Results: The proposed KG-SCCA method is able to model two types of prior knowledge: one as a group structure (e.g., linkage disequilibrium blocks among SNPs) and the other as a network structure (e.g., gene co-expression network among brain regions). The new model incorporates these prior structures by introducing new regularization terms to encourage weight similarity between grouped or connected features. A new algorithm is designed to solve the KG-SCCA model without imposing the independence constraint on the input features. We demonstrate the effectiveness of our algorithm with both synthetic and real data. For real data, using an Alzheimer's disease (AD) cohort, we examine the imaging genetic associations between all SNPs in the *APOE* gene (i.e., top AD gene) and amyloid deposition measures among cortical regions (i.e., a major AD hallmark). In comparison with a widely used SCCA implementation, our KG-SCCA algorithm produces not only improved cross-validation performances but also biologically meaningful results.

Availability: Software is freely available upon request.

Contact: shenli@iu.edu

PP37 - Large-scale automated identification of mouse brain cells in confocal light sheet microscopy images

Paolo Frasconi[1], Ludovico Silvestri[2], Paolo Soda[3], Roberto Cortini[1], Francesco Pavone[2] and Giulio Iannello[3]

[1]Department of Information Engineering (DINFO), Università di Firenze, Italy. [2]European Laboratory for Nonlinear Spectroscopy (LENS), Università di Firenze, Italy. [3]Integrated Research Centre, Università Campus Bio-Medico di Roma, Italy.

ABSTRACT

Motivation: Recently, confocal light sheet microscopy has enabled high-throughput acquisition of whole mouse brain 3D images at the micron scale resolution. This poses the unprecedented challenge of creating accurate digital maps of the whole set of cells in a brain.

Results: We introduce a fast and scalable algorithm for fully automated cell identification. We obtained the whole digital map of Purkinje cells in mouse cerebellum consisting of a set of 3D cell center coordinates. The method is very accurate and we estimated an F_1 measure of 0.96 using 56 representative volumes, totaling 1.09 GVoxel and containing 4,138 manually annotated soma centers.

Availability and implementation: Source code and its documentation are available at http://bcfind.dinfo.unifi.it/. The whole pipeline of methods is implemented in Python and makes use of ylearn2 (Goodfellow et al., 2013) and modified parts of Scikitlearn (Pedregosa et al., 2011). Brain images are available

on request.
Contact: paolo.frasconi@unifi.it
Supplementary information: Coordinates of predicted soma centers of a whole mouse cerebellum and additional figures.

Highlight Talk: *HP09 - Novel developments in computational clinical breath analysis and biomarker detection.*

Anne-Christin Hauschild[1], Jörg Ingo Baumbach[2] and Jan Baumbach[3]

[1]Max Planck Institute for Informatics, Saarbrücken, Germany. [2]Faculty Applied Chemistry, Reutlingen University, Germany. [3]University of Southern Denmark, Denmark.

ABSTRACT

The volatolom is the sum of volatile organic compounds that are emitted by all living cells and tissues. We seek to non-invasively “sniff” biomarker molecules that are predictive for the biomedical fate of individual patients. This promises great hope to move the therapeutic windows to earlier stages of disease progression. While portable devices for breathomics measurement exist, we face the traditional biomarker research barrier: a lack of robustness hinders translation to the world outside laboratories. To move from biomarker discovery to validation, from separability to predictability, we have developed several bioinformatics methods for computational breath analysis, which have the potential to redefine non-invasive biomedical decision making by rapid and cheap matching of decisive medical patterns in exhaled air. We aim to provide a supplementary diagnostic tool complementing classic urine, blood and tissue samples. The presentation will review the state of the art, highlight existing challenges and introduce new data mining methods for identifying breathomics biomarkers.

Publications:

Hauschild AC, Kopczynski D, D'Addario M, Baumbach JI, Rahmann S, Baumbach J. Peak detection method evaluation for ion mobility spectrometry by using machine learning approaches. Metabolites. 2013 Apr 16;3(2):277-93.

Eckel SP, Baumbach J, Hauschild AC. On the importance of statistics in breath analysis--hope or curse? J Breath Res. 2014 Mar;8(1):012001.

Maurer F, Hauschild AC, Eisinger K, Baumbach J, Mayor A and Baumbach JI. MIMA - a software for analyte identification in MCC/IMS chromatograms by mapping accompanying GC/MS measurements. Int. J. Ion Mobil. Spec. 2014 Apr:17:95–101.

Smolinska A, Hauschild AC, Fijten RR, Dallinga JW, Baumbach J, van Schooten FJ. Current breathomics-a review on data pre-processing techniques and machine learning in metabolomics breath analysis. J Breath Res. 2014 Jun;8(2):027105.

Contact: jan.baumbach@imada.sdu.dk

Wed3 (Area H): Text Mining for Computational Biology

Chairs: To be announced

PP38 - Extracting patterns of database and software usage from the bioinformatics literature

Geraint Duck[1], Goran Nenadic[1,2], Andy Brass[1,3], David Robertson[3] and Robert Stevens[1]

[1]School of Computer Science, University of Manchester, UK. [2]Manchester Institute of Biotechnology, University of Manchester, UK. [3]Computational and Evolutionary Biology, Faculty of Life Sciences, University of Manchester, UK.

ABSTRACT

Motivation: As a natural consequence of being a computer-based discipline, bioinformatics has a strong focus on database and software development, but the volume and variety of resources are growing at unprecedented rates. An audit of database and software usage patterns could help provide an overview of developments in bioinformatics and community common practice, and comparing the links between resources through time could demonstrate both the persistence of existing software and the emergence of new tools.
Results: We study the connections between bioinformatics resources and construct networks of database and software usage patterns, based on resource co-occurrence, that correspond to snapshots of common practice in the bioinformatics community. We apply our approach to pairings of phylogenetics software reported in the literature, and argue that these could provide a stepping-stone into the identification of scientific best practice.
Availability: The extracted resource data, the scripts used for network generation and the resulting networks are available at: http://bionerds.sourceforge.net/networks/
Contact: robert.stevens@manchester.ac.uk

Highlight Talk: *HP10 - Text mining technologies for database curation*

Fabio Rinaldi[1], Simon Clematide[1], Simon Hafner[1], Gerold Schneider[1], Gintare Grigonyte[2], Martin Romacker[3] and Therese Vachon[4]

[1]University of Zurich, Switzerland. [2]University of Stockholm, Sweden. [3]F. Hoffmann-LaRoche, Switzerland. [4]Novartis, Switzerland.

ABSTRACT

Although human curation for life science databases offers the best guarantee of high quality annotations, it suffers from severe bottlenecks which have long been recognized in the curation community. The most pressing problem is that of efficiency of the process: it is impossible for human curators to keep up with the growing pace of publication. Text mining technologies, coupled with advanced user interfaces, offer the potential to partially alleviate this bottleneck. We survey the results of several recent competitive evaluations of text mining technologies, discuss how text mining systems can be integrated in a curation workflow, and illustrate our approach to assisted curation, which has been tested in collaboration with major databases.

Publication :

Rinaldi F, Clematide S, Hafner S, Schneider G, Grigonyte G, Romacker M, Vachon T. Using the OntoGene pipeline for the triage task of BioCreative 2012. Database (Oxford). 2013 Feb 9;2013:bas053.

Contact: fabio@ontogene.org

Wed4 (Areas A and E): RNA prediction

Chair: Yann Ponty

PP39 - CRISPRstrand: Predicting repeat orientations to determine the crRNA-encoding strand at CRISPR loci

Omer S. Alkhnbashi[1], Fabrizio Costa[1], Shiraz A. Shah[2], Roger A. Garrett[2], Sita J. Saunders[1] and Rolf Backofen[1,3]

[1]Bioinformatics Group, Department of Computer Science, University of Freiburg, Georges-Köhler-Allee 106, 79110 Freiburg, Germany, [2]Archaea Centre, Department of Biology, University of Copenhagen, Ole Maaloes Vej 5, DK2200 Copenhagen, Denmark, [3]BIOSS Centre for Biological Signalling Studies, Cluster of Excellence, University of Freiburg, Germany.

ABSTRACT

Motivation: The discovery of CRISPR-Cas systems almost 20 years ago rapidly changed our perception of the bacterial and archaeal immune systems. CRISPR loci consist of several repetitive DNA sequences called repeats, inter-spaced by stretches of variable length sequences called spacers. This CRISPR array is transcribed and processed into multiple mature RNA species (crRNAs). A single crRNA is integrated into an interference complex, together with CRISPR-associated (Cas) proteins, to bind and degrade invading nucleic acids. Although existing bioinformatics tools can recognize CRISPR loci by their characteristic repeat-spacer architecture, they generally output CRISPR arrays of ambiguous orientation and thus do not determine the strand from which crRNAs are processed. Knowledge of the correct orientation is crucial for many tasks, including the classification of CRISPR conservation, the detection of leader regions, the identification of target sites (protospacers) on invading genetic elements, and the characterization of protospacer-adjacent motifs (PAMs).

Results: We present a fast and accurate tool to determine the crRNA-encoding strand at CRISPR loci by predicting the correct orientation of repeats based on an advanced machine learning approach. Both the repeat sequence and mutation information were encoded and processed by an efficient graph kernel to learn higher order correlations. The model was trained and tested on curated data comprising more than 4,500 CRISPRs and yielded a remarkable performance of 0.95 AUC ROC (area under the curve of the receiver operator characteristic). In addition, we show that accurate orientation information greatly improved detection of conserved repeat sequence families and structure motifs. We integrated CRISPRstrand predictions into our CRISPRmap web server of CRISPR conservation and updated the latter to version 2.0.

Availability: CRISPRmap and CRISPRstrand are available at http://rna.informatik.uni-freiburg.de/CRISPRmap

Contact: backofen@informatik.uni-freiburg.de

PP40 - Towards a piRNA prediction using multiple kernel fusion and support vector machine

Jocelyn Brayet[1,2], Farida Zehraoui[1], Laurence Jeanson-Leh[2], David Israeli[2] and Fariza Tahi[1]

[1]IBISC, UEVE/Genopole, IBGBI, 23 bv. de France, 91000 Evry, France. [2]Genethon, 1, bis rue de

l'Internationale, 91002 Evry Cedex, France.

ABSTRACT

Motivation: Piwi interacting RNA (piRNA) is the most recently discovered and the least investigated class of AGO/Piwi protein interacting small non-coding RNAs. PiRNAs are mostly known to be involved in protecting the genome from invasive transposable elements. But recent discoveries suggest their involvement in the pathophysiology of diseases, such as cancer. Their identification is therefore an important task, and computational methods are needed. However, the lack of conserved piRNA sequences and structural elements makes this identification very challenging and difficult.

Results: In the present study, we propose a new modular and extensible machine learning method based on multiple kernels and a support vector machine (SVM) classifier for piRNA identification. Very few piRNA features are known to date. The use of a multiple kernels approach allows editing, adding or removing piRNA features that can be heterogeneous in a modular manner according to their relevance in a given species. Our algorithm is based on a combination of the previously identified features (sequence features (k-mer motifs and a uridine at the first position) and piRNAs cluster feature) and a new telomere/centromere vicinity feature. These features are heterogeneous and the kernels allow to unify their representation. The proposed algorithm, named piRPred, gives very promising results on Drosophila and Human data and outscores previously published piRNA identification algorithms.

Availability: piRPred is freely available to non-commercial users on our Web server EvryRNA: http://EvryRNA.ibisc.univ-evry.fr

Contact: tahi@ibisc.univ-evry.fr