BeNeLux Bioinformatics Conference 2013

December 9-10 - Brussels, Belgium

"Bioinformatics
at all scales of life..."

http://bbc2013.ibsquare.be

INDEX

Cover page photo courtesy of Nikolaj Tatti.

Welcome to the 8th edition of the BeNeLux Bioinformatics Conference!

This year, BBC 2013 takes place in Brussels (Belgium) on December 9-10, 2013.

The conference is organised by the *Université Libre de Bruxelles* and the *Vrije Universiteit Brussel*, which have recently joined their efforts in the bioinformatics field with the creation of the Interuniversity Institute of Bioinformatics in Brussels $(IB)^2$.

The BeNeLux Bioinformatics Conference is an annual conference that started in 2005, bringing together researchers from the Benelux and beyond to strengthen national and international connections in all fields of bioinformatics. The exchange will focus on advances in bioinformatics and applications in the areas of methodology development, biotechnology, and health.

The theme of this year's conference is: **"Bioinformatics at all scales of life"**

The bioinformatics research field is inherently interdisciplinary and combines biology with different approaches from computer science, mathematics, ...

The aim of this discipline is to tackle biological problems by means of data analysis and modeling. These biological problems arise at all scales of life: at the genome level, at the protein level, at the cellular level, or at the level of a complete organism, like plants, or mammals.

The topics that will be covered include, but are not limited to: the study and modeling of biological networks (genes regulatory networks, protein-protein interaction networks, metabolic networks, ...), structural bioinformatics (modeling and study of the structures of biomolecules and their interactions), and the analysis of high-throughput sequencing data or micro-array expression data.

We hope you all have a successful conference, with plenty of networking and sharing of information that can stimulate ideas and collaborations.

The BBC 2013 organising committee

CONFERENCE VENUE

Salle Dupréel

Address:
Université Libre de Bruxelles (ULB) Solbosch campus – Building S
Avenue Jeanne – Johannalaan 44 1050 Brussels (Ixelles – Elsene) Belgium
google maps coordinates; 50.814707,4.38023

Solbosch Campus, Université Libre de Bruxelles (ULB)

IMPORTANT REMARKS ABOUT THE VENUE

- The oral presentations will be in the Salle Dupréel lecture room, which is at the first floor level of the venue building.
- Coffee, lunch and poster sessions will be in the hall outside the Salle Dupréel lecture room and will be spread over the ground and first floor levels of the hall. We want to urge delegates to make optimal use of the available space and spread over the two floors of the hall.
- Please be advised that the ground floor and first floor hall are public areas in which we have to allow passage of those wanting to reach any of the 15 floors of the building. Therefore, it is best not to leave valuable items unattended.
- As a service to those coming straight to the venue with their luggage, we will aim to organise a luggage drop- off at the venue to the best of our ability. However, using the luggage drop-off implies you accepting that any damage, loss or theft of items in the luggage drop-off remains at your own risk.
- Upon registration, you will be provided your personal BBC2013 congress badge. Make sure to wear your congress badge at all times. Entrance to the Salle Dupréel lecture room and participation to the coffee and lunch breaks and to the reception will be restricted to those wearing a BBC2013 congress badge.
- Your congress badge should also be worn as a proof of reservation for the congress dinner.

BBC 2013 SPONSORS

- **PLATINUM SPONSORS**

visit.brussels

fnrs LA LIBERTÉ DE CHERCHER

B2

gsk GlaxoSmithKline Vaccines

- **GOLD SPONSORS**

INSILICO GENOMICS AT YOUR FINGERTIPS

- **SILVER SPONSORS**

HistoGeneX

- **BRONZE SPONSORS**

IBM

iSCB INTERNATIONAL SOCIETY FOR COMPUTATIONAL BIOLOGY

The BeNeLux Bioinformatics Conference 2013 is also an affiliated conference of the International Society of Computational Biology.

Extend the Life of your Poster The BeNeLux Bioinformatics Conference 2013, in cooperation with Faculty of 1000, invites all presenters to deposit any of their poster/oral presentation(s) slides into the established open access poster repository F1000 Posters, to allow those who could not make the meeting have the opportunity to see your novel work. To deposit your research, simply go to http://f1000.com/posters/depositor and upload your file to maximize the value of your conference presentations.

F1000Posters

COMMITTEES BBC 2013

- **ORGANISING COMMITTEE**
  - **Elisa Cilia**, Université Libre de Bruxelles
  - **Kevin D'hoe**, Vrije Universiteit Brussel, VIB
  - **Yves Dehouck**, Université Libre de Bruxelles
  - **Pierre Geurts**, Université de Liège
  - **Dimitri Gilis**, Université Libre de Bruxelles
  - **Tom Lenaerts**, Université Libre de Bruxelles and Vrije Universiteit Brussel
  - **Yvan Saeys**, VIB, University of Ghent
  - **Wim Vranken**, Vrije Universiteit Brussel

- **SCIENTIFIC PROGRAM COMMITTEE**
  - **Francisco Azuaje**, CRP-Santé, Luxembourg
  - **Gianluca Bontempi**, Université Libre de Bruxelles
  - **Didier Croes**, UZ Brussel
  - **Bernard De Baets**, University of Ghent
  - **Vincent Detours**, Université Libre de Bruxelles
  - **Johan den Dunnen**, Leiden University Medical Center
  - **Chris Evelo**, Maastricht University
  - **Peter Horvatovich**, University of Groningen
  - **Martijn Huynen**, CMBI, UMC St. Radboud, Nijmegen
  - **Gunnar Klau**, Centrum Wiskunde & Informatica (CWI), Amsterdam
  - **Kris Laukens**, University of Antwerp
  - **Steven Maere**, VIB, University of Ghent
  - **Lennart Martens**, VIB, University of Ghent
  - **Perry Moerland**, Academic Medical Centre (AMC), Amsterdam
  - **Pieters Monsieurs**, SCK-CEN, Mol
  - **Yves Moreau**, Katholieke Universiteit Leuven
  - **Jan-Peter Nap**, Wageningen University & Research Centre
  - **Jeroen Raes**, Vrije Universiteit Brussel
  - **Dick de Ridder**, Delft University of Technology
  - **Jeroen de Ridder**, Delft University of Technology
  - **Peter de Rijk**, University of Antwerp
  - **Marianne Rooman**, Université Libre de Bruxelles
  - **Pierre Rouzé**, VIB, University of Ghent
  - **Thomas Sauter**, University of Luxembourg
  - **Roland Siezen**, UMC St. Radboud, Nijmegen
  - **Nicolas Simonis**, Université Libre de Bruxelles
  - **Guillaume Smits**, Hôpital Universitaire des Enfants Reine Fabiola (HUDERF), Bruxelles
  - **Berend Snel**, Utrecht University
  - **Jacques van Helden**, Université d'Aix-Marseille (AMU)
  - **Sacha van Hijum**, UMC St. Radboud / NIZO Food Research
  - **Klaas Vandepoele**, VIB, University of Ghent
  - **Gert Vriend**, CMBI, UMC St. Radboud, Nijmegen

PROGRAM

MONDAY DECEMBER 9, 2013

From 9:00 am: Registration

Session 1. Chair: Yvan Saeys

9:30 - 10:00	Opening, including $(IB)^2$ speech
10:00 - 10:30	Alejandro Sifrim - KU Leuven ***eXtasy: variant prioritization by genomic data fusion (#73)***
10:30 - 10:45	Aalt-Jan van Dijk - Wageningen University and Research Centre ***Connecting phenotypes and traits to biological processes and molecular functions (#77)***
10:45 - 11:00	Zeynep Kalender Atak - KU Leuven ***Comprehensive analysis of transcriptome variation uncovers known and novel driver events in T-cell acute lymphoblastic leukemia (#80)***

11:00 - 11:30: Coffee/Tea break, with biscuits

Session 2. Chair: Chris Evelo

11:30 - 12:00	Tim De Meyer - Ghent University ***Applications of Large-Scale Genome-Wide DNA Methylation Profiling (#42)***
12:00 - 12:15	Giulia Gonnelli - VIB, Ghent University ***Logistic regression for the classification of PSMS: a simple method for a complex problem (#4)***
12:15 - 12:30	Raphaël Marée - University of Liège ***Large-scale bioimage analysis using web services and machine learning (#70)***

12:30 - 14:00: Lunch and company presentations

13:30 - 13:45: David Weiss – InSilico. *InSilico DB: all the genomics data you create, ready to analyse how and by who you want.*
13:45 - 14:00: Yves Sucaet – Histogenex. *The pragmatic bioinformatician.*

Session 3. Chair: Tom Lenaerts

14:00 - 15:00	**Alfonso Valencia** - Spanish National Cancer Research Centre (CNIO) ***Co-Evolution Based Methods in the Prediction and Analysis of Protein Interaction Networks***
15:00 - 15:30	Francisco Azuaje - CRP-Santé, Luxembourg ***Improving the detection of biologically meaningful clusters in protein interaction networks through integrated functional analysis (#20)***
15:30 - 15:45	Rita Pancsa - VIB, Vrije Universiteit Brussel ***DynaMine: from protein sequence to dynamics and disorder (#10)***
15:45 - 16:00	Daniele Raimondi - Interuniversity Institute of Bioinformatics Brussels, ULB-VUB ***Prediction of protein residues contacts with Deep Learning and Direct Information methods (#3)***

16:00 - 16:30: Coffee/Tea break, with biscuits

Session 4. Chair: Wim Vranken

16:30 - 17:30	**Anna Tramontano** - Sapienza University of Rome ***The computational analysis of biomolecular interactions***
17:30 - 18:30	Poster commercials 1. Patrice Godard - Thomson-Reuters, UCB - *In silico drug repurposing in Parkinson's disease (#86)* 2. Georgios Dalkas - Université Libre de Bruxelles - *Computational analysis of antigen-antibody compared to other protein-protein interactions (#15)* 3. Wout Bittremieux - University of Antwerp - *jqcML: A Java API for quality control for mass spectrometry experiments (#91)* 4. Sandra Steyaert - Ghent University - *SNP-guided identification of monoallelic DNA-methylation events from enrichment-based sequencing data (#43)* 5. Elvis Ndah - Ghent University - *Protein identification based on ribosome targeted mRNA fragments (#29)* 6. Hana Imrichová - KU Leuven - *Integrative analysis of regulatory tracks for the identification of direct TF-target interactions (#21)* 7. Dries De Maeyer - KU Leuven - *Mechanistic interpretation of gene lists using interaction networks (#31)* 8. Şule Yılmaz - VIB, Ghent University - *Comparing fragmentation spectra from two parasitic worm species to discover unique peptides (#62)* 9. Vanessa Vermeirssen - VIB, Ghent University - *Integrating gene regulatory network inference solutions for the abiotic stress response in Arabidopsis thaliana (#67)* 10. Raf Winand - KU Leuven, iMinds - *Prediction accuracy for deleterious and disease causing mutations in healthy individuals (#74)*

18:30 - 22:00: Poster session, walking dinner, and beer tasting

TUESDAY DECEMBER 10, 2013

Session 5. Chair: Francisco Azuaje

9:30 - 9:45	Opening
9:45 - 10:15	Mark de Been - University Medical Centre Utrecht ***Whole-genome sequence-based identification of epidemic plasmids spreading Extended-Spectrum Beta-Lactamase genes among E. coli from different hosts (#49)***
10:15 - 10:30	Koen Illeghems - Vrije Universiteit Brussel ***Towards a software-independent taxonomic profiling method of microbial metagenomes (#58)***
10:30 - 10:45	Bas E. Dutilh - CMBI, NCMLS, UMC Radboud ***Comparative metagenomics by cross-assembly (#60)***

10:45 - 11:15: Coffee/Tea break, with biscuits

Session 6. Chair: Gert Vriend

11:15 - 11:45	Fredrick M. Mobegi - UMC Radboud, CMBI ***Rapid identification of potential antimicrobial drug targets (#46)***
11:45 - 12:00	Jan Fostier - Ghent University, iMinds ***Comparative Motif Discovery In The Cloud (#106)***
12:00 - 12:15	Dmitry Svetlichnyy - KU Leuven ***Prediction of transcriptional targets using advanced enhancer model (#22)***

12:15 - 13:45: Lunch and software demos

13:00 - 13:15: Lars Eijssen - Maastricht University. *DBXP: investigating the future of integrative bioinformatics research infrastructures in Europe (#95)*
13:15 - 13:30: Rekin's Janky - KU Leuven. *iRegulon: Detecting master regulators and cis-regulatory interactions in human cancer related gene networks (#81)*
13:30 - 13:45: Fotis Georgatos - University of Luxembourg. *Facilitating computational biology and bioinformatics on HPC systems using EasyBuild (#93)*

Session 7. Chair: Sacha van Hijum

13:45 - 14:45	**Jim Haseloff** - University of Cambridge ***Engineering simple plant systems***
14:45 - 15:15	Riet De Smet - VIB, Ghent University ***Convergent gene loss following gene and genome duplications creates single-copy families in flowering plants (#66)***

15:15 - 15:45: Coffee/Tea break, with biscuits

Session 8. Chair: Kris Laukens

15:45 - 16:00	Robrecht Cannoodt - Ghent University ***Network inference by integrating biclustering and feature selection (#33)***
16:00 - 16:15	Catharina Olsen - Université Libre de Bruxelles ***Data-driven validation of gene regulatory networks using knock-down data (#37)***
16:15 - 16:30	Mafalda Galhardo - Université du Luxembourg ***Integrated analysis of transcript level regulation of metabolism reveals disease relevant nodes of the human metabolic network (#88)***
16:30 - 16:45	Diana Hendrickx - Maastricht University ***Comparison of methods for pattern recognition in toxicogenomics time series (#27)***
16:45 - 17:15	Closing and Prizes

17.15 - 18:15: Farewell drinks and snacks

INVITED SPEAKERS

- **JIM HASELOFF**, University of Cambridge

Jim Haseloff is a plant biologist working at the Department of Plant Sciences, University of Cambridge. His scientific interests are focused on the engineering of plant morphogenesis, using microscopy, molecular genetic, computational and synthetic biology techniques. Prior to joining the Department of Plant Sciences, Jim served as group leader at MRC Laboratory of Molecular Biology in Cambridge and his group developed advanced imaging techniques and modified fluorescent proteins for efficient use in plants. Before this, Jim was a research fellow at Harvard Medical School, working on trans-splicing ribozymes. He has also worked at the CSIRO Division of Plant Industry, Canberra, and developed methods for the design of the first synthetic RNA enzymes with novel substrate specificities. Jim is deeply involved with teaching Synthetic Biology at the University of Cambridge, and is very interested in its wider potential as a tool for engineering biological systems and underpinning sustainable technologies.

Engineering simple plant systems

Synthetic Biology has great potential as a tool for the engineering of multicellular organisms. (1) The greatest diversity of cell types and biochemical specialisation is found in multicellular systems, (2) the molecular basis of cell fate determination is increasingly well understood, and (3) it is feasible to consider creating new tissues or organs with specialized biosynthetic or storage functions by remodelling the distribution of existing cell types. Of all multicellular systems, plants are the obvious first target for this type of approach. Plants possess indeterminate and modular body plans, have a wide spectrum of biosynthetic activities, can be genetically manipulated, and are widely used in crop systems for production of biomass, food, polymers, drugs and fuels. Current GM crops generally possess new traits conferred by single genes, and expression results in the production of a new metabolic or regulatory activity within the context of normal development. However, cultivated plant varieties often have enlarged flowers, fruit organs or seed, and are morphologically very different from their wild-type ancestors. The next generation of transgenic crops will contain small gene networks that confer self-organizing properties, with the ability to reshape patterns of plant metabolism and growth, and the prospect of producing neomorphic structures suited to bio production. We have developed a battery of computational, imaging and genetic tools to allow clear visualisation of individual cells inside living plant tissues and have the means to reprogram them. These techniques are well suited to study of simple experimental systems such as the lower plant Marchantia polymorpha and surrogate microbial populations. These types of simple systems are becoming increasingly important to explore the next generation of genetic circuits with self-organising properties.

http://www.haseloff-lab.org
http://www.synbio.org.uk

- **ANNA TRAMONTANO**, Sapienza University of Rome

Anna Tramontano was trained as a physicist but she soon became fascinated by the complexity of biology and by the promises of computational biology. After a post-doctoral period at UCSF, she joined the Biocomputing Programme of the EMBL in Heidelberg. In 1990 she moved back to Italy to work in the Merck Research Laboratories near Rome. In 2001, she returned to the academic world as a Chair Professor of Biochemistry in "La Sapienza" University in Rome where she continues to pursue her scientific interests on protein structure prediction and analysis and on genomic and post-genomic data interpretation in the Department of Physics. She is a member of the Scientific Council of the ERC, of the European Molecular Biology Organization, the Scientific Council of Institute Pasteur - Fondazione Cenci Bolognetti, and the organizing Committee of the Critical Assessment of Techniques for Protein Structure Prediction (CASP) initiative. She is a member of the Advisory Board of the SIB in Basel, of the CRG in Barcelona, of the CNB in Madrid, of the MPI for Molecular Genetics in Berlin, of the IIMCB in Warsaw and has been a member of the EMBL Scientific Adivsory Committee and of the EBI Advisory Committee. She is Associate Editor of Bioinformatics, Proteins, PLoS One and Current Opinion in Structural Biology.

She was awarded the prize for Natural Sciences of the Italian Government, the “Marotta Prize” of the Italian National Academy of Science and the Minerva Prize for Scientific Research, the KAUST Investigator Award and has published four books (Bioinformatica - Zanichelli; The ten most wanted solutions in Protein Bioinformatics - CRC Press; Protein Structure Prediction - Wiley; Introduction to Bioinformatics - CRC Press).

The computational analysis of biomolecular interactions

The combination of experimental and computational approaches can provide invaluable information on the function of biological systems. The computational approach is usually based on empirical methods for predicting the three-dimensional structure and the function of a protein, as well as its interactions with both macromolecules and smaller compounds. These methods, even if still approximate, can be instrumental for the development of effective rational strategies for experiments such as studies of disease related mutations, site directed mutagenesis, or structure based drug design. I will describe some of the methods that we developed to this end, including a method to dissect and predict antibody-antigen interactions. I will also show some examples of their effectiveness in providing relevant information about systems of biomedical interest.

- **ALFONSO VALENCIA**, Spanish National Cancer Research Center

Alfonso Valencia is a biologist with formal training in population genetics and biophysics which he received from the Universidad Complutense de Madrid. He was awarded his PhD in 1988 at the Universidad Autónoma de Madrid. He was a Visiting Scientist at the American Red Cross Laboratory in 1987 and from 1989-1994 was a Postdoctoral Fellow at the laboratory of C. Sander at the European Molecular Biology Laboratory (EMBL), Heidelberg, Germany. In 1994 Alfonso Valencia set up the Protein Design Group at the Centro Nacional de Biotecnología, Consejo Superior de Investigaciones Científicas (CSIC) in Madrid where he was appointed as Research Professor in 2005. He is a Member of the European Molecular Biology Organisation (EMBO), Founder and former Vice President of the International Society for Computational Biology where he has been Chair of the Systems Biology and/or Text Mining Tracks of the main Computational Biology Annual Conference (ISMB) since 2003. He was honoured as ISCB-Fellow in 2010. Alfonso Valencia serves on the Scientific Advisory Board of the European Molecular Biology Laboratory; the Swiss Institute for Bioinformatics, Biozentrum, Basel; the INTERPRO database; the Spanish Grant Evaluation Agency (ANEP); as well as the Steering Committee of the European Science Foundation Programme on Functional Genomics (2006-2011). Alfonso Valencia is Co-Executive Editor of Bionformatics, serves on the Editorial Board of EMBO Journal and EMBO Reports, among others. He is the Director of the Spanish National Bioinformatics Institute (INB).

Co-Evolution Based Methods in the Prediction and Analysis of Protein Interaction Networks

Co-evolution based methods are potentially able to increase the coverage and accuracy of the information provided by high-throughput protein-protein interaction methods (1). Given the importance of protein interaction networks to contextualize functional information in genome analysis pipelines, it is relevant to explore the possibilities that this methodology open. Recent publications in the area of protein folding have revitalized the interest in the use of co-evolution based methods for the prediction of protein interactions. In this presentation, I will first review the general panorama of co-evolution based methods with particular emphasis in those applied to the prediction of protein interactions (2). In the second part of the talk, I will present a new co-evolution based approach to the prediction of protein interaction networks in a large set of bacterial species (3). The radical innovation of this new approach is that it permits for the first time the prediction of specie specific interactions instead of the general type of predictions for protein families provided by other approaches.

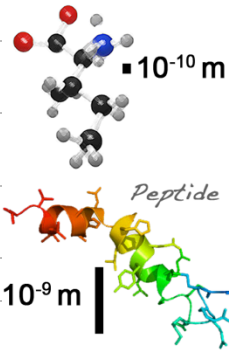
Funding MICROME. EU Framework Programme 7 Collaborative Project. Grant Agreement Number 222886-2 Spanish Government grant BIO2007-66855

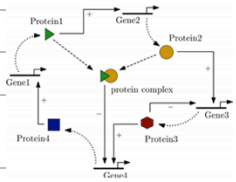
References

(1) de Juan D, Pazos F, Valencia A. (2013) Emerging methods in protein co-evolution. Nat Rev Genet. 14:249-261.
(2) Mosca E, Pons T, Ceol A, Valencia A, Aloy P. Towards a detailed atlas of protein-protein interactions. Curr Opin Struct Biol 2013 (PMID 23896349).
(3) de Juan D, de la Torre V, Valencia A (2013) in preparation.

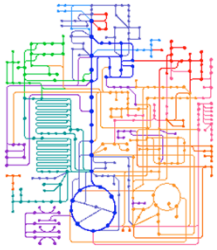
LIST OF ABSTRACTS: BIOINFORMATICS AT ALL SCALES OF LIFE

Oral Presentation *Software Demo* *Poster Commercial* *Poster Presentation*



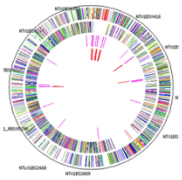


Gene regulation Networks

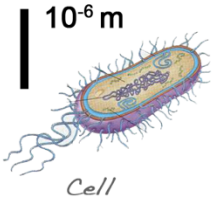


Metabolism

Genome

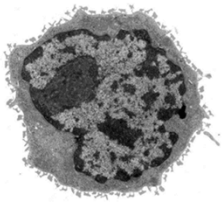


Sequencing

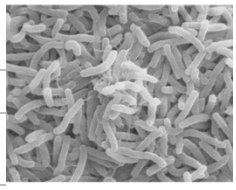


Cell

10^{-5} m



Cell population



Metagenome

61 ORCAE: online resource for community annotation of eukaryotes.

62 Comparing fragmentation spectra from two parasitic worm species to discover unique peptides.

63 Identifying losses and expansions of selected genes families in incomplete genomic datasets.

64 Unravelling the genetic basis of *Fusarium* sugarbeet wilt disease.

65 Sequence based genotyping: applications for plant breeding.

66 Convergent gene loss following gene and genome duplications creates single-copy families in flowering plants.

67 Integrating gene regulatory network inference solutions for the abiotic stress response in *Arabidopsis thaliana*.

68 Visual analysis of spermatozoa, oocytes and early embryonic transcripts.

69 Tensor decomposition for data reduction in mass spectrometry imaging.

70 Large-scale bioimage analysis using web services and machine learning.

71 Accuracy of an automated vessel counting algorithm in four different tumor types.

72 Guided exploration of mass spectrometry imaging data through integration with anatomical information.

73 eXtasy: variant prioritization by genomic data fusion.

74 Prediction accuracy for deleterious and disease causing mutations in healthy individuals.

75 Convert your favourite protein modelling program into a mutation predictor: "MODICT".

76 Unravelling the regulatory mechanisms behind inter-genic cardiac quantitative trait loci through systems genetics approaches.

77 Connecting phenotypes and traits to biological processes and molecular functions.

78 A class representative model for pure parsimony haplotyping under uncertain data.

79 Classifying the progression of ductal carcinoma from single-cell sampled data: a case study.

80 Comprehensive analysis of transcriptome variation uncovers known and novel driver events in T-cell acute lymphoblastic leukemia.

81 Detecting master regulators and cis-regulatory interactions in human cancer related gene networks.

82 Long non-coding RNAs in lung cancer: comparison of microarray and RNA-seq techniques.

83 Extracting signatures from high-dimensional unbalanced biological data: the cases of DNA methylation and LNCRNA in breast cancer.

84 BELLEROPHON: a hybrid method for detecting interchromosomal rearrangements at base pair resolution using next-generation sequencing data.

85 A human-specific endogenous retroviral insert serves as an enhancer for the schizophrenia-linked gene PRODH.

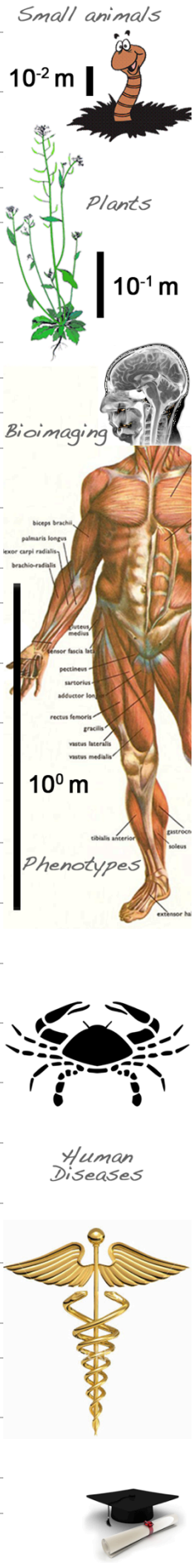
86 In silico drug repurposing in Parkinson's disease.

87 Assessment of reference networks for pathway analysis and mechanistic interpretation of disease data.

88 Integrated analysis of transcript level regulation of metabolism reveals disease relevant nodes of the human metabolic network.

89 Biomedical text mining for disease gene discovery.

90 Bioinformatics and systems biology masters: bridging the gap between heterogeneous student backgrounds.



91 jqcML: A Java API for quality control for mass spectrometry experiments.

92 Identifying interesting frequent patterns in complex biological data with MIME.

93 Facilitating computational biology and bioinformatics on HPC systems using EasyBuild.

94 NGS logistics: data infrastructure for efficient analysis of NGS sequence variants.

95 DBXP: investigating the future of integrative bioinformatics research infrastructures in Europe.

96 Data integration & stewardship centre: tackling the big data challenge in life science research.

97 Bioinformatics @ DSM biotechnology centre

98 PacBio: single molecule sequencing to improve the Norway Spruce genome annotation.

99 A random forests based breast cancer diagnosis tool using circulating miRNA expression.

100 Regression with enriched random forest.

101 Transposable element annotation using relational random forests.

102 Small decision models: capitalizing on feature selection.

103 Predicting tryptic cleavage from proteomics data using decision tree ensembles.

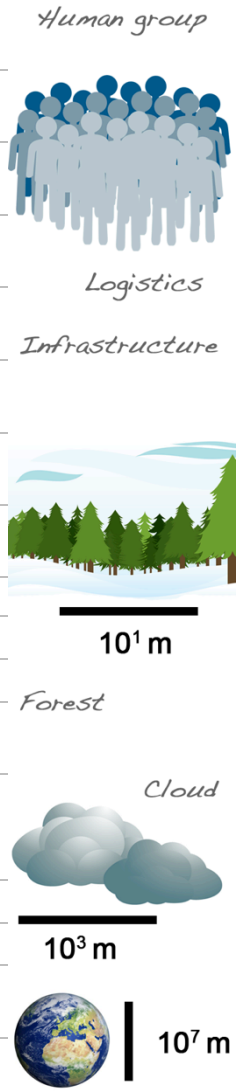
104 Promoting a functional and comparative understanding of the conifer genome- implementing applied aspects for more productive and adapted forests.

105 NMR_REDO: Large-scale recalculation of NMR structures in the cloud.

106 Comparative motif discovery in the cloud.

107 Large-scale ancestral route reconstruction of the Luxembourgian HIV cohort in international context.

108 Species interactions in the world's oceans.



ARTEFACTS IN THE REFINEMENT OF ISOLEUCINE

Karen RM Berntsen[*], *Erik Stens & Gert Vriend.*
CMBI, Radboud University Nijmegen Medical Centre, 6525 GA 26-28, The Netherlands.
[*]*k.berntsen@student.science.ru.nl*

χ1 and χ2 of isoleucines in X-ray protein structures were observed to be a function of resolution, secondary structure, and refinement software used. At low resolution, the average torsion angle values for the nine rotameric states differs between refinement software that uses molecular dynamics-like energy terms (CNS[1]) and software that does not use these terms (REFMAC[2]). Detailing the standard torsion angles used in refinement software can improve the refinement of protein structures.

INTRODUCTION

The bond length and bond angle parameters of Engh and Huber[3] are used in nearly all macromolecular software available today. Other authors have been working on torsion angle parameters and the use of them in refinement software.

The target values for the parameters such as bond lengths and angles are used similar as data in crystallographic refinement. Consequently it can be expected that their final values will be closer to reality at high resolution and closer to the target values at low resolution. If the target values can be improved, especially low resolution models will improve.

MacArthur and Thornton[4] already realized that the average values of observed side chain torsion angles are resolution dependent. Touw and Vriend[5] showed later on that the ideal τ angles depend on secondary structure and amino acid type, and the values observed in PDB files additionally depend on the resolution and the refinement software. We study if these factors similarly influence the observed isoleucine side chain torsion angles and conclude that the determination of the optimal χ1-χ2 values is a highly complicated task that is open for improvements.

METHODS

All X-ray PDB files released before 01-02-2013 were stored in a relational database, augmented with torsion angles, secondary structure[6] and residual WHAT_CHECK[7] quality parameters. The distribution of each of the nine rotameric states of the isoleucine side chain is analysed in χ1-χ2 space for different datasets.

An isoleucine in a poly-alanine helix was periodically rotated around its χ1 and χ2. For each conformation the total energy (in Joule/mol) and the atom experiencing the greatest force were stored. This latter atom colours the ln(energy) contours in the χ1-χ2-energy plot. The same procedure is performed on an isoleucine in a β-hairpin.

The rotameric state of 940.338 isoleucines with atomic B-factors < 80 in PDB and PDB-REDO[8] were compared. From the 5,3% differences the net flux between rotameric states was calculated.

RESULTS & DISCUSSION

The rotamers of isoleucine can be divided into nine highly unevenly populated sections, which is mainly caused by strain due to interactions between the side chain atoms and the local backbone atoms. The average value of χ1 and χ2 depends on resolution, secondary structure, backbone torsion angles and refinement software used.

We observed that conformations that are most populated in experimentally determined protein structures correspond to areas where the energy (according to the YASARA-Nova[9] force field) is optimal. The location and depth of the local minima in the energy plots correspond well enough with the maxima in the frequency plots to explain the observed frequency differences in the two times (α-helix and β-strand) 9 sections. See Figure 1.

The PDB still does not hold enough data to make adequately detailed subset selections needed to answer each question that is left. In future studies we additionally want to use culled datasets that contain only one copy of any series of similar structures solved by the same group with the same software.

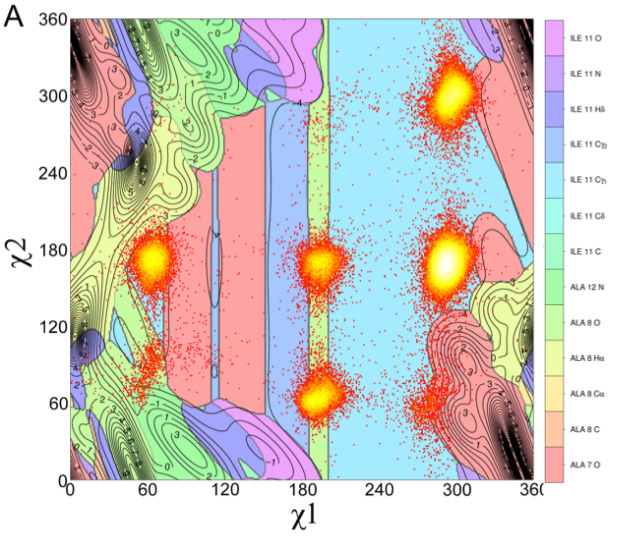


***FIGURE 1**. Figure 1 shows in colour for 360[2] conformations of isoleucine in an α-helix which atom experiences the largest force using the YASARA-Nova[9] force field. Gray lines separate coloured areas. Black lines connect conformations with equal energy.*

REFERENCES

1. Brünger, *et al*. Acta Cryst. D54, 905–921 (1998).
2. Murshudov GN et al. *Acta Cryst*. **D53**, 240–255 (1997).
3. Engh RA & Huber R *Acta Cryst.* **A47**, 392-400 (1991).
4. MacArthur, MW & Thornton JM. *Acta Cryst*. **D55**,994-1004 (1999).
5. Touw WG & Vriend G. *Acta Cryst*. **D66**, 1341-1350 (2010).
6. Kabsch W & Sander C. *Biopolymers* **22**, 2577-2637 (1983).
7. Hooft RW et al. *Nature* **381**, 272 (1996).
8. Joosten RP & Vriend G. *Science* **317**, 195-196 (2007).
9. Krieger et al. *Proteins: Struc. Func.* (2002).

IDENTIFYING CHOLESTASIS-CAUSING DRUG COMPOUNDS: A VALIDATED LIGAND-BASED PHARMACOPHORE MODEL

Susanne M.A. Hermans[1,2], Rick Greupink[2], Marieke Schreurs[2], Jeroen J.M.W van den Heuvel[2], Jan B. Koenderink[2], Frans G.M. Russel[2] and Tina Ritschel[1].*

*[1] Computational Discovery and Design (CDD) Group, Centre for Molecular and Biomolecular Informatics (CMBI), Radboud University Medical Centre, Nijmegen, The Netherlands. [2] Pharmacology & Toxicology Group, Nijmegen Centre for Molecular Life Sciences (NCMLS), Radboud University Medical Centre, Nijmegen, The Netherlands. [3] Netherlands eScience Center, Amsterdam, The Netherlands. *T.Ritschel@umcn.nl*

A ligand-based pharmacophore model, a 3D-model, fixing the intra molecular distances between important molecular properties of the ligands, is built to detect whether molecules can inhibit BSEP. The model may be used as an *in silico* screening tool to help eliminate potentially harmful drug candidates at an early stage in drug development, saving time and money and improving drug safety.

INTRODUCTION

Many drugs on the market have the potential to induce cholestasis. Cholestasis is a blockage of bile flow, leading to hepatotoxicity which can result in jaundice and ultimately liver failure. Drug-induced cholestasis is often the results of an unexpected, off-target interaction with the bile salt export pump (BSEP), the key membrane transporter responsible for the transport of bile acids from hepatocytes into the bile[1].

METHODS

Inhibitors were identified using transport uptake experiments on vesicles transfused with human BSEP. The compounds that showed significant BSEP inhibition (>50%) were aligned with endogenous substrates to create a pharmacophore model describing the molecular features necessary to bind to BSEP. The model was validated using a set of 60 drug-like compounds. A pharmacophore screening of the CoCoCo database was performed to select putative BSEP inhibitors[3, 4]. An RMSD ranking and a virtual inspection were performed to select six diverse compounds for further *in vitro* validation of this pharmacophore model[2].

RESULTS & DISCUSSION

Five out of 32 compounds showed BSEP inhibition (>50%) in the vesicle transport uptake experiments. The five inhibitors (grey) were aligned with two bile acids (black) to identify common molecular features used to create a pharmacophore model (Figure 1).

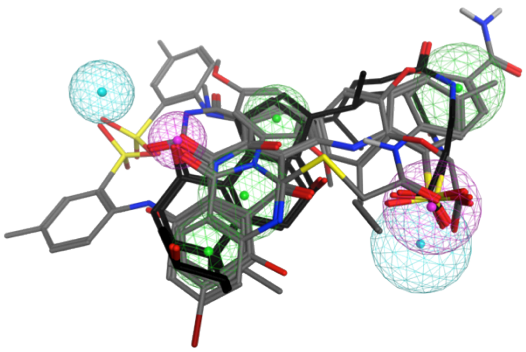


***FIGURE 1**. The final BSEP pharmacophore model[4].*

The final model consisted of four hydrophobic/aromatic features (green) and two hydrogen bond acceptor/anionic features (pink) combined with two hydrogen bond acceptor projection features (cyan) and 26 size exclusion volumes (not shown).

The model recognized 77% of the inhibitors and 15% of the non-inhibitors of the 60 drug-like compounds used for validation.

A virtual screening of the CoCoCo database resulted in the identification of six structurally diverse putative BSEP inhibitors[3]. Experimental validation showed that all 6 compounds show BSEP inhibition 4 of which with inhibition >50%.

CONCLUSION

In order to identify BSEP inhibitors a BSEP pharmacophore model was created using experimentally identified inhibitors. The model was experimentally validated with a set of 60 drug-like compounds and six compounds identified from the CoCoCo database. This model improves the identification of BSEP inhibitors significantly. The BSEP pharmacophore may therefore aid in the *in silico* identification of potential cholestasis-inducing drug candidates at the early stages of drug development.

REFERENCES

1. Alrefai WA, Gill RK: Bile acid transporters: structure, function, regulation and pathophysiological implications. *Pharm Res* 24,1803-1823 (2007).

2. Dror O, Schneidman-Duhovny D, Inbar Y, Nussinov R, Wolfson HJ: Novel approach for efficient pharmacophore-based virtual screening: method and applications. *J Chem Inf Model* 49, 2333-2343 (2009).

3. Del Rio A, Barbosa AJ, Caporuscio F, Mangiatordi GF: CoCoCo: a free suite of multiconformational chemical databases for high-throughput virtual screening purposes (http://www.cococo-database.it). *Mol Biosyst* 6, 2122-2128 (2010).

4. Molecular Operating Environment (MOE), 2011.10. *Chemical Computing Group Inc, 1010 Sherbooke St West, Suite #910, Montreal, QC, Canada, H3A 2R7,* 2011.

PREDICTION OF PROTEIN RESIDUES CONTACTS WITH DEEP LEARNING AND DIRECT INFORMATION METHODS

Marcin J. Skwark[1,2,3], Daniele Raimondi[1,4,] and Arne Elofsson[1,2].*

*Department of Biochemistry and Biophysics, Stockholm University, 10691 Stockholm, Sweden[1]; Science for Life Laboratory, Box 1031, 17121 Solna, Sweden [2] ; Department of Information and Computer Science, Aalto University, PO Box 15400, FI-00076 Aalto, Finland[3]; Interuniversity Institute of Bioinformatics Brussels,ULB-VUB, La Plaine Campus, Triomflaan, BC building, 6th floor, CP 263 1050 Brussels, Belgium[4] . * eddiewrc@alice.it*

Recently, several new contact prediction methods have been published and are clearly superior to earlier methods when it comes to predicting contacts in proteins. They use large sets of multiple aligned sequences and assume that correlations between columns in these alignments can be the results of residue interactions and thus clues of residues spatial proximity in the native structure. To further improve the quality of these predictions we developed a Deep Learning (DL) architecture able to abstract some typical inter-residue relationships among neighbouring residue pairs, namely it learns to visually recognize frequent Secondary Structure (SS) patterns in the Contact Maps (CM). DL is able to correctly re-evaluate some nonsense predictions and, at the same time, to cluster residues pairs around the regions of the contact maps with real biological and structural significance, providing a significant improvement of the precision of the overall CM.

INTRODUCTION

The DL architecture is a further improvement of PconsC[1], a CM meta-predictor that combines predictions from two direct information methods, PSICOV[2] and plmDCA[3], calculated from two alignment methods, HHblits and jackHmmer, at four different e-value thresholds, obtaining an improvement of the predictive performances with respect to the single methods on which it is based. The aim of the DL predictor is to further enhance the quality of these predictions.

METHODS

DL is a sub-field of Machine Learning in which several layers of representation are learnt, obtaining a hierarchical representation of the data that mimics the concept of *concepts*. The DL architecture presented here is a feed-forward 4-layer stack of Random Forest Classifiers. Each layer $L_{k>1}$ takes into account both a *standard* feature vector and a subset of the predictions made by the previous layer L_{k-1}, natively including the structure underlying the Contact Prediction (CP) problem into its learning horizons by considering the predicted neighbourhood of each residues pair. We also considered some commonly used features (predicted SS, predicted RSA and Sequence Profile).

RESULTS & DISCUSSION

CMs prediction is an intrinsically non-local problem; for this reason we developed a DL architecture able to perform structured predictions by taking into consideration the significant amount of information underlying the CP problem instead of simply considering each residue pair independent from the others (in [4] has been shown how contacts in the native structure can hardly involve a single pair of residues). Thanks to this *neighbourhood awareness*, each layer of the DL architecture evaluates the contacting state of each residue pair conditioning its likelihood to be in contact on the basis of its structural context, predicted by the previous layer.

DL is useful when you need to abstract higher level concepts from lower level ones; the ultimate goal in this case is to add a sufficient number of abstraction layers in order to obtain a DL architecture able to *understand* the concept of CM as a whole, predicting them entirely and respecting their biological, structural and physical meaningfulness. This kind of architectures should became able to predict CMs that really resemble *evolution made* proteins, integrating a sufficient amount of knowledge in order to tell apart, among all the sparse matrices, which ones belong to the fairly small subset that represents biologically meaningful protein structure.

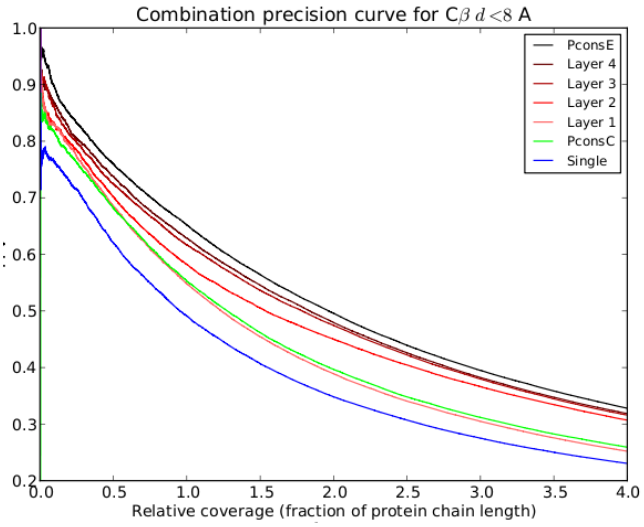


FIGURE 1. *PPV curves showing the prediction performances. In blue is shown the best single method, in green PconsC and in shades of red the improvement provided by each layer of DL. The black curve shows the final results obtained by DL with extended set of features.*

REFERENCES

1. Marcin J. Skwark et al., doi:10.1093/bioinformatics/btt259
2. Jones et al., Bioinformatics, **28**, 184–190 (2012).
3. Ekeberg et al., Phys Rev E Stat Nonlin Soft Matter Phys, **87**, 012707 (2013).
4. Di Lena et al., (2012) doi:10.1093/bioinformatics/bts4755.

LOGISTIC REGRESSION FOR THE CLASSIFICATION OF PSMS: A SIMPLE METHOD FOR A COMPLEX PROBLEM

Giulia Gonnelli[1, 2], *Michiel Stock*[3], *Jan Verwaeren*[3], *Sven Degroeve*[1, 2], *Bernard De Baets*[3] & *Lennart Martens*[1, 2*]

[1]*Department of Medical Protein Research, VIB, B-9000 Ghent, Belgium;* [2]*Department of Biochemistry, VIB, B-900 Ghent, Belgium;* [3]*Dept. of Mathematical Modelling, Statistics and Bioinformatics, Ghent, Belgium. *lennart.martens@ugent.be*

The growing interest in proteogenomics or metaproteomics studies has revealed serious limitations in the application of the "decoy" database paradigm, used to separate correct peptide identifications from incorrect ones in traditional shotgun proteomics. Indeed, as the sequence database increases dramatically in size, and may start to contain very many different but highly similar peptides, the number of potential incorrect peptide-to-spectrum matches (PSMs) increases as the decoy approach fails to discriminate between correct and incorrect hits. Moreover, one has to take into account the dramatic increase in computing time necessary when processing whole genome-size databases.

INTRODUCTION

Traditionally, in mass spectrometry-based proteomics, target-decoy approaches are used to establish the false discovery rate (FDR), the ratio of incorrect to correct Peptide-to-Spectrum Matches (PSMs) at a certain cutoff. In these approaches, the experimental spectra are also searched against nonsense, "decoy" databases (often created by reversing or shuffling the sequences in the original database), and the number of hits in such a database is taken to be an estimate of the number of false hits obtained by searching the original database, so FDR = #decoys / #originals. However, this approach is not always optimal, especially for Proteogenomics[1] where whole genomes are searched and the vast increase of high-scoring false hits makes it extremely difficult to separate true hits from false ones, and consequently to estimate an FDR. Additionally, popular and reliable software tools like Percolator[2] need to be retrained each time for each dataset, a process that consumes impractically large amounts of computing time in proteogenomics due to the combined size of the original and the decoy database. In this work, we therefore developed a new FDR estimation method that no longer relies on the traditional "decoy" databases. Instead, we trained a linear model once on a collection of heterogeneous data to classify correct and incorrect PSMs. We then used this classifier to look for a dataset-independent score threshold that can provide correct PSMs at a known FDR.

METHODS

We used simple and fast L1-regularized logistic regression to build a binary classifier. The classifier was trained only once on two thousands PSMs chosen at random, from different organisms, and processed using different settings, retrieved from the ms-lims repository[3]. We used Mascot ranks to label correct and incorrect PSMs: rank 1 was taken to be a correct match, while ranks 2, 3, 4, and 5 are chosen to model incorrect PSMs. We built ten different models according to the rank used to model incorrect PSMS.

RESULTS & DISCUSSION

All models perform quite well, and interestingly, the model trained on rank1 hits together with decoy hits shows the lowest performance when tested on datasets in which ranks lower than rank1 are used to model incorrect hits, indicating that decoys are not very good at modeling incorrect PSMs (Figure 1). Additionally, in order to validate the reliability of the predictions made by the classifier, we searched for a general score cutoff to separate correct from incorrect PSMs. Preliminary results on yeast data showed that a score threshold could be identified reinforcing the robustness of the method.

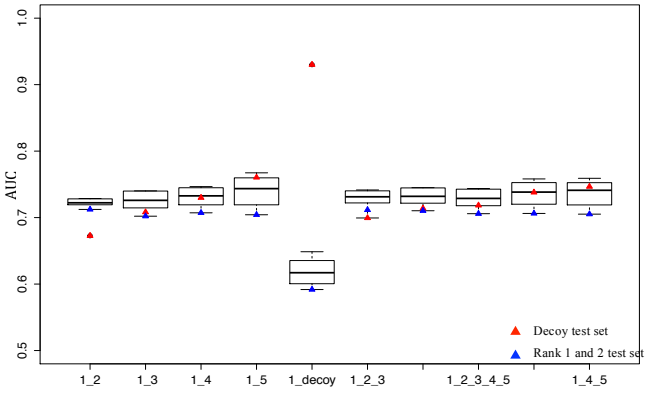


FIGURE 1. Boxplots of the AUC for ten different models trained and tested on ten different datasets. *All models perform similarly. The model trained on rank 1 and decoy shows the highest performance when tested on the rank 1 and decoy test set but shows the lowest performance when tested on the other dataset, indicating that decoys are able to model random hits accurately, but not higher-scoring incorrect ones. Labels for the different models are reported on the x-axis.*

We thus present a powerful method to confidently and quickly classify PSMs and which does not require retraining for each data set.

REFERENCES

1. Renuse S, et al. *Proteogenomics Proteomics*. **11**, 620-30 (2011).
2. Käll L. et al. *Nat Methods*.**4**, 923-5 (2007).
3. Helsens K. et al.. *Proteomics* **10**, 1261-4 (2010).

FOCUS ON RELATIVELY HYDROPHILIC PEPTIDES FOR TARGETED PROTEOMICS

Nicolas Housset[1,2*], *Sven Degroeve*[1,2], *Lennart Martens*[1,2].
Dept. of Biochemistery, Universiteit Gent[1] *; Dept. of Biomedical Protein Research, VIB*[2].
[*]*nicolas.housset@ugent.be*

Targeted Proteomics aims at selecting the best transitions in order to identify and eventually quantify proteins. Proteins yielding many tryptic peptides, we want to target the best peptides. Observing that the precision of retention time prediction gets lower as we move further into the gradient, we investigate the possibility of protein identification using only the more hydrophilic peptides.

INTRODUCTION

Liquid chromatography (LC) coupled to mass spectrometry (MS) is one of the main techniques used in proteomics. LC releases peptides gradually to the mass spectrometer: we can record the moment a peptide is detected after release as the retention time. In general, the higher the hydrophobicity of a peptide, the longer its retention time will be. Targeted proteomics needs this retention time information for selected reaction monitoring (SRM) scheduling: the mass spectrometer will be tuned to specifically look for one peptide in a window around its expected retention time.

Precision of retention time prediction is thus critical: a smaller confidence interval means a smaller window and additional possibilities for the mass spectrometer to target peptides.

METHODS

A number of proteins from different model organisms is studied. The increasing complexity of those organisms is connected to the increasing complexity of their proteomes.

Performing an *in silico* trypsin digestion, we then answer this question: for each organism, what proportion of their proteins has at least one hydrophilic peptide, two, three?

We consider a peptide to be hydrophilic if it is predicted to elute in the first half of the gradient. We use relative gradients because the total length is subject to change across laboratories.

For retention time prediction, we use ELUDE[1] as trained on data stored in MS-LIMS[2].

RESULTS & DISCUSSION

The metric used to assess the performance of ELUDE is the size (relative to the gradient) of the retention time window so that 95% of the peptides elute in this time window.

We observe (Figure 1) that the precision of the retention time predictor decreases as we move further in the gradient.

A single protein yields many tryptic peptides, it is thus best to target the peptides where the retention time prediction is better.

In the ideal case of the best theoretical retention time predictor, we explore the possibility of protein identifications using only the more hydrophilic peptides.

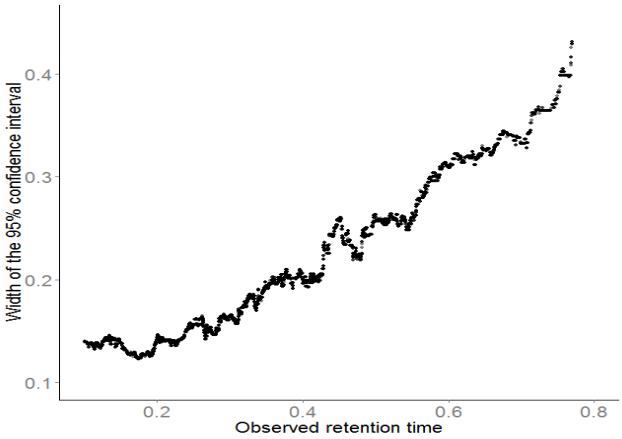


FIGURE 1. Width of the 95% confidence interval. *We observe a steady increase in the width as we move further in the gradient.*

A number of proteins from different organisms is studied: for simple organisms, we expect proteins to be less numerous and shorter. The probability of finding hydrophilic enough peptides may then be lower. For more complex organisms, proteins are longer and we can expect to find hydrophilic peptides in the majority of the proteins.

A liquid chromatography experiment can be quite long, thus reducing the length, even slightly, of this part will lead to substantial time-saving when running many samples.

Since retention time confidence intervals of hydrophilic peptides are lower, we can increase the number of peptides targeted in the first half of the gradient.

REFERENCES

1. Moruz L., Staes A., Foster J. M., Hatzou M., Timmerman E., Martens L., Käll L. Proteomics 2012, 12 1151-1159.
2. Helsens K, Colaert N, Barsnes H, Muth T, Flikka K, Staes A, Timmerman E, Wortelkamp S, Sickmann A, Vandekerckhove J, Gevaert K, Martens L. Proteomics. 2010 Mar;10(6):1261-4.

CODING REGIONS SUBJECT TO MULTIPLE CONSTRAINTS TEND TO ENCODE INTRINSICALLY DISORDERED PROTEIN SEGMENTS

Mauricio Macossay-Castillo[1], Simone Kosol[1,2], Peter Tompa[1,2] and Rita Pancsa[1,2]*

*[1]Structural Biology Brussels, VUB, Belgium. [2]Department of Structural Biology, VIB, Belgium. *rpancsa@vub.ac.be*

Certain genomic regions code for multiple, overlapping functions, which can be detected by analyzing the levels and patterns of their evolutionary conservation. Thousands of such potentially multi-functional elements, namely synonymous constraint elements (SCEs), were recently discovered in human coding exons. We hypothesized that the protein segments encoded by such elements might better tolerate the multiple constraints stemming from the increased functional demands if they are of lower levels of structural constraints (i.e. disordered protein segments). SCEs were mapped onto human proteins and a variety of structure-prediction methods were applied to study the resulting segments. SCE-encoded protein segments are significantly enriched in structural disorder and low sequence complexity, and depleted in secondary structure elements and domain annotations. These observations are further validated on a set of protein segments encoded by experimentally confirmed exonic splice regulatory sites. Our results provide evidence that gene regions of increased functional demands tend to colocalize with structural disorder in the encoded proteins. However, this coincidence does not explain which came first.

INTRODUCTION

Recently, around 10000 SCEs were identified within human protein-coding exons[1]. SCEs are coding regions, in which synonymous mutation rates are extremely low, thus indicating additional sequence constraints beyond protein coding. The additional functions are mostly regulatory sites involved in translation initiation and transcript splicing. In such multi-functional gene regions the amino acid spectrum of the encoded protein segments are restricted by the second functionality. Due to this reason, we hypothesized that these protein regions have a reduced capability to form secondary structure elements and that they are likely enriched in structural disorder.

METHODS

We have used the Perl API of Ensembl to find the corresponding protein segments for the published SCE genomic locations. We used different measures to describe the structural properties of the resulting segments: 1) the fraction of predicted disordered residues, 2) the fraction of low-complexity residues, 3) the fraction of predicted secondary structure elements, and 4) the fraction of residues in Pfam entities. Reference protein segment sets were created by randomly picking an equally long segment from human SCE-containing proteins for each identified protein segment. The structural properties of these were determined similarly. All human, experimentally verified, exonic splicing factor binding sites (SFBSs) were collected from the SpliceAid-F database[2] and filtered for redundancy. The corresponding protein segments were also subjected to similar structural analyses.

RESULTS & DISCUSSION

Our results clearly support that the SCE-encoded protein regions are biased towards lower structural constraints. The relatively large size of SCE datasets also allowed for finding correlations between the strength of the above mentioned structural biases and the window size of SCE detection. As a further proof of the principle, the SFBSs overlapping protein segments showed similar tendencies. The structural effects of overlapping functionalities are demonstrated on HOXA2 (Figure 1).

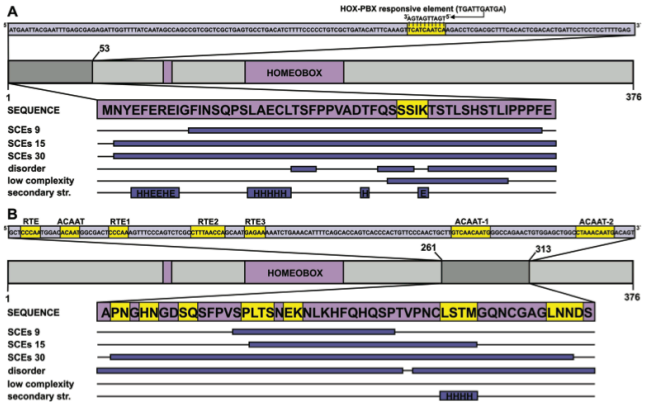


***FIGURE 1**. Protein structural effects of secondary functions in coding regions: the HOXA2 gene. The Hox-A2 protein is represented by a light grey bar, with domains colored purple and SCE-encoded regions marked by dark grey. The CDS is shown above with the regions of multi-functionality highlighted in yellow for the A) N-terminal and B) the enhancer rich C-terminal SCE. The structural properties of the corresponding protein segments are indicated below.*

Our results reflect that the level of complexity encoded by a genomic region of given length is limited and in case of multiple competing functions this limitation necessitates compromising. Since regulatory functions at DNA or RNA level are fulfilled by short stretches of nucleotides, their information content cannot be reduced. Proteins, however, have many positions which are not crucial for their functions and structural integrity, and are thus rather free to change. This is particularly true for regions of structural disorder and low sequence complexity[3]. In accord, we found that genomic regions of multiple functionalities are more likely to overlap with disordered protein regions, which suggests that the encoded complexity is rather evenly distributed in the coding regions of genomes.

REFERENCES

1. Lin MF et al. *Genome research* **21**, 1916-1928 (2011).
2. Giluietti M et al. *Nucleic Acids Res* **41**, D125-131 (2013).
3. Brown CJ et al. J Mol Evol **55**, 104-110 (2002).

SPATIALLY COHESIVE AMINO ACIDS AND THEIR ROLE IN PROTEIN MOLECULAR STRUCTURES

Pieter Meysman[1,2,*], *Cheng Zhou*[1], *Boris Cule*[1], *Bart Goethals*[1] & *Kris Laukens*[1,2].
ADReM, Department of Mathematics and Computer Science, University of Antwerp, Antwerp, Belgium[1]*; Biomedical Informatics Research Center Antwerp (biomina), University of Antwerp / Antwerp University Hospital, Edegem, Belgium*[2].
[*]*pieter.meysman@uantwerpen.be*

In this work, we examine the role that spatially co-occurring amino acids play in protein molecular structures. To this end, we apply a novel algorithm that is able to mine cohesive patterns in multi-dimensional structure data to a variety of protein molecular structures. While several found patterns represent conserved domains in proteins, as could be expected, other patterns seem to occur outside of any known protein domain. Comparison to the literature and to molecular energy models reveals that many of these patterns are important contributors to the stability of the protein structure.

INTRODUCTION

Proteins in living cells exist as complex three dimensional macromolecules consisting primarily of a single chain of amino acids. The complexity of the mechanisms that drive the folding of these proteins and the evolutionary constraints that impact this biological process, remain an important research question. In this work, we set out to identify which amino acids frequently co-occur in protein molecular structures and examine the role that they play within the structure.

METHODS

For the identification of spatially co-occurring amino acids, we apply a novel structural itemset miner on a large data set of protein three-dimensional structures, as extracted from the PDB database[1]. This miner is based on the concept of cohesion[2], which is a metric to consistently calculate the proximity of items in a pattern that has been extended here to work for multidimensional structural data. The advantage of cohesion is that it removes the need of a strict cut-off on individual instances of the pattern, which might result in missing interesting patterns. This method thus allows discovery of interesting patterns within a multidimensional spatial structure by combining the cohesion and frequency of the pattern[3]. In this specific case, the patterns to be found consist of sets of amino acids that co-occur in close proximity with high frequency in a set of proteins.

RESULTS & DISCUSSION

We applied the cohesive itemset miner to two types of datasets. The first consisted of protein structures that shared a certain common feature, either a common domain, such as a winged helix DNA-binding domain, or a common function, such as kinase activity. The second type consisted of the entire non-redundant PDB database of almost 30 000 protein structures. Within these datasets, we consistently found patterns of three or four amino acids that spatially co-occur with a high frequency. Despite not including any sort of constraint, many of the found patterns consist of amino acids that are very distant in the protein sequence. Further, many of these patterns describe patterns that match amino acids in different regions of the protein structure and not only in the functional domains, see Figure 1.

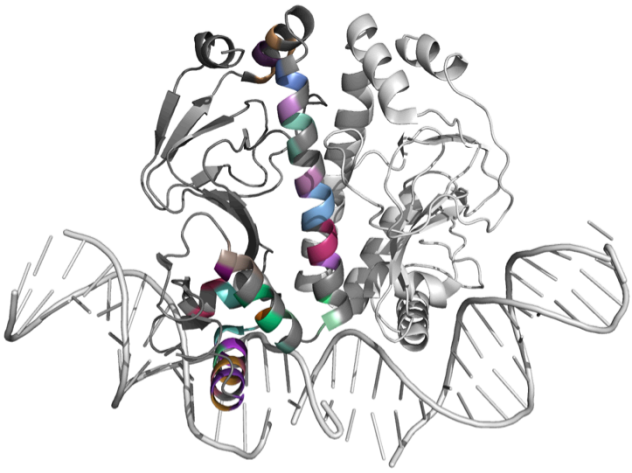


***FIGURE 1**. The molecular structure of the E. coli CRP transcription factor bound to a DNA molecule. The residues matching frequent patterns found for the winged helix proteins are highlighted in color.*

To examine the role of these amino acid patterns on the protein structure, we compared them to external data sources, such as the conserved domains from Pfam database[4] and the literature. This analysis revealed that several of the frequent patterns consistently occur in protein regions outside of known conserved domains and several combinations of residues are known to play an important role in the overall stability of the protein tertiary or quaternary structure. To further investigate this finding, we calculated the relative energy contributions of each residue using the FoldX software[5]. This analysis revealed that the energy contribution of the residues matching the frequent patterns do indeed significantly differ from those that do not.

REFERENCES

1. Kouranov, A. *et al. Nucleic acids research* **34,** D302–5 (2006).
2. Cule, B., Goethals, B. & Robardet, C. in *Proceedings of the SIAM International Conference on Data Mining (SDM)* (2009).
3. Zhou C. et al. *Proceedings of the 12th International Workshop on Data Mining in Bioinformatics* 42–50 (2013).
4. Punta, M. *et al. Nucleic acids research* **40,** D290–301 (2012).
5. Schymkowitz, JWH. *et al.. PNAS (USA)* **102,** 10147–52 (2005).

IN SILICO STABILITY ANALYSIS METHOD APPLIED TO BOVINE SEMINAL RIBONUCLEASE

Marie De Laet[*], *Yves Dehouck, Dimitri Gilis, Marianne Rooman.*
Dept. of BioModeling, BioInformatics & BioProcesses, Université Libre de Bruxelles. [*]*madelaet@ulb.ac.be*

The majority of native contacts in natural proteins contribute favourably to the structure's stability but some regions are found to be locally destabilizing. Recent findings support the idea that residues belonging to functional sites or initiation sites of conformational changes have been optimized during evolution to fulfill a specific function, but not necessarily to ensure stability. A program that detects protein regions that are particularly optimal or non-optimal with respect to thermodynamic stability is presented here, together with the results of it application to the model system bovine seminal ribonuclease.

INTRODUCTION

Improving predictive methods for functional annotation of protein structures is still a challenge. The limitation of computational methods[1,2] often comes from the need to know related structures of the target protein to infer characteristics of their catalytic or binding sites. Other techniques are based on sequence information only[3], Both approaches are limited by the fact that similar proteins do not always share same biological function. Concurrently, structural studies suggest that key functional residues are located in regions that contribute unfavourably to the protein thermodynamic stability[4,5].

We present here an *in silico* tool that measures the energetic contributions of residues to the overall thermodynamic stability of the protein and that detects regions of particularly high or low stability. Our approach is applied to an interesting and illustrative study case, the bovine seminal ribonuclease (BS RNase). This protein, member of the well-know pancreatic ribonuclease superfamily, catalyzes the degradation of RNA strands and undergoes 3D domain swapping in physiological conditions. The available high quality crystal structures of different conformations of BS RNase have been analysed to examine the relationship with it functional sites and regions involved in conformational changes.

METHODS

Database-derived statistical potentials were used to compute the contribution of each residue to the overall protein stability. They are derived from observed frequencies of association of specific sequence and structure elements[6]. The stability is estimated using three types of database-derived statistical potentials: a distance potential, a backbone torsion angle potential and a solvent accessibility potential. These potentials allow the identification of stability peculiarities with respect to the tertiary interactions, the local structure or the core/surface pattern, respectively.

A residue that has an energetic contribution significantly lower than the average contribution computed in a reference set of proteins is detected as a structural weakness. Conversely, a structural robustness corresponds to a residue that brings substantial contribution to the overall protein stability. Our program groups structural weaknesses and robustnesses that are in contact into tridimensional stability patches.

RESULTS & DISCUSSION

Stability patches detected in bovine seminal ribonuclease structures are located in the vicinity of functional sites. In particular, a structural weakness is identified in the catalytic site, which supports previous observations that functional residues can have a cost in regard to the protein stability. We also detected other stability patches that overlap the binding sites.

BS RNase exists in an equilibrium mixture of two dimeric forms. One of these is the non-swapped dimer that converts into the swapped form through the breaking of interactions in the "closed interface" and the exchange of identical parts between the two chains. The new interactions between the monomers that do not exist in the monomeric form are part of the "open interface". These structural interfaces are detected separately by distinct stability patches. Interestingly, the closed interface contains a majority of structural weaknesses mostly detected by the torsion potential. This suggests that non-optimal regions to the protein stability can locally unfold to allow a 3D domain swapping process. In the unswapped structure, the stability patch including residues of the open interface loses a structural weakness in the swapped counterpart, which is thus stabilized upon swapping. This observation is consistent with experimental findings[7].

These results are promising for further analysis of functional features and conformational changes and for the design of a predictive tool.

REFERENCES

1. Polacco B.J. and Babbitt P.C. Automated discovery of 3D motifs for protein function annotation. Bioinformatics 22, 723-730 (2006).
2. Zhang C. and Kim S. Overview of structural genomics: from structure to function. Curr Opin Chem Biol 7, 28-32 (2003).
3. Gerlt JA. and Babbitt PC. Can sequence determine function? Genome Biol. 1, 5 (2000).
4. Dessailly B.H., Lensink M.F. & Wodak S.J. Relating destabilizing regions to known functional sites in proteins. BMC bioinformatics, 8, 141 (2007).
5. Ferreiro D.U., Hegler J.A., Komives E.A. & Wolynes P.G. On the role of frustration in the energy landscapes of allosteric proteins. Proc Nat Ac Sci, 108(9), 3499–503 (2011).
6. Dehouck Y., Gilis D., Rooman M. A New generation of Statistical Potentials for Proteins. Biophysical J. 90, 4010-4017 (2006).
7. Ercole C., Avitabile F., Del Vecchio P., Crescenzi O, Tancredi T, Picone D. Role of he hinge peptide and the intersubnit interface in the swapping of N-termini in dimeri bovine seminal RNase. Eur J Biochem 270, 4729-4735 (2003).

MAPPING INTRA-PROTEIN COMMUNICATION –THE FYN SH2 SNAP-LOCK MECHANISM

Radu Huculeci[1,2], Elisa Cilia[3], Lieven Buts[1,2], Klaartje Houben[4], Nico van Nuland[1,2] & Tom Lenaerts[3,5].*

*Structural Biology Brussels, Vrije Universiteit Brussel[1]; Department of Structural Biology, VIB[2]; MLG, Département d'Informatique, Université Libre de Bruxelles[3]; Bijvoet Center for Biomolecular Research, Utrecht University[4]; AI-lab, Vakgroep Computerwetenschappen, Vrije Universiteit Brussel[5]. *tlenaert@ulb.ac.be*

Efficient communication, mediated by proteins, is an essential property of all biological systems. Recently, different computational methods have been proposed to identify the amino-acids involved in transmitting the information throughout a protein structure, focussing especially on cases where no large structural changes can be observed in the backbone. Here we report an original method that combines experimental and computational methods to identify intra-protein communication patterns, using the human Fyn SH2 domain as a model. These results reveal for the first time the intra-molecular information exchange that may lie at the heart of the activity regulation of the Src tyrosine kinases, of which Fyn is a leading example.

INTRODUCTION

Human Fyn, a member of Src family kinases, performs essential roles in cellular signalling and its aberrant functionality has been associated with a wide variety of diseases, including cancer. Fyn is a multi-domain protein, composed of SH3, SH2 and catalytic domains. The kinase activity is down-regulated through the coordinated binding of its SH2 and SH3 domains in a so called "snap-lock" configuration. This mechanism assumes that an efficient intra-protein communication occurs within the SH2 domain, as a first step of the kinase inactivation process. Our work provides a hypothesis on how this communication may actually work.

METHODS

In order to decipher the intra-protein communication pathways that are at play in the Fyn SH2 domain, both its free state and its complex with the natural inhibitory peptide were analyzed. The approach proposed here combines deuterium-based NMR methyl relaxation (DNMR) experiments[1] with predictions produced by an *in silico* framework[2] to identify the set of residues involved in propagating the binding information through the structure.

Previous NMR work[3] clearly demonstrated that DNMR experiments[1] provide a powerful tool to study internal dynamics with atomic resolution. To identify the residues involved in information transmission through the protein domain, we have performed an extensive NMR dynamics analysis of Fyn SH2, both at the backbone and side-chain levels. The DNMR experiments provide internal dynamical data only on the methyl bearing residues, and thus offer only a partial image of the long-range dynamical effects at the protein level upon peptide binding. By making use of our predictive approach[2], we complemented the experimental results to include also non-methyl bearing residues, while ensuring an optimal overlap with the experimental results. This *in silico* framework uses Monte-Carlo sampling and Shannon's information theory to quantify the conformation coupling between the residue side-chains[2].

RESULTS & DISCUSSION

The combined experimental and predictive approach provides a method to identify and quantify information transfer within protein domains. Applying this approach to human Fyn SH2, the DNMR measurements showed different degrees of freedom for some isolated methyl side-chains. These effects were experienced by methyl groups located both at short and long (more than 18 Å) distances from the peptide-binding interface. The network of most prominent changes, which is inferred by extending the results with predictions (see Figure 1), provides a direct connection between the SH2 peptide binding site and the linker connecting the SH2 and SH3 domains. This linker is known to rigidify and maintain the inactive form of the kinase. Our results provide a model that can explain how the dephosphorylation of the tail of Fyn may trigger a signal through the SH2 domain, rendering the linker flexible and allowing the Fyn kinase to become active. As such this work offers a hypothesis for the activation process of the Src-like kinases, requiring further work to unravel the roles of the identified core residues.

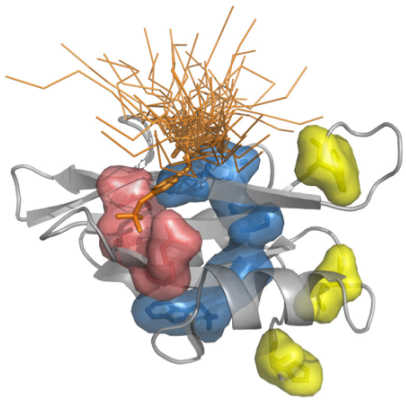


***FIGURE 1**. Three different clusters of highly dynamically coupled residues (highlighted in three different code colours) through the Fyn SH2 domain induced by the natural inhibitory peptide binding. The residues are mapped on our Fyn SH2 bound state (the NMR conformer closest to the mean structure - grey), while the peptide is shown in all the conformations from the NMR ensemble (orange).*

REFERENCES

1. Houben KL et al. *Biophys J* **93**, 2830-44 (2007).
2. Cilia E et al. *PLOS Comp. Biol.* **8**, e1002794 (2012).
3. Fuentes EJ et al. *J Mol Biol.* **335**, 1105-15 (2004).

DYNAMINE: FROM PROTEIN SEQUENCE TO DYNAMICS AND DISORDER

Elisa Cilia[1,5], Rita Pancsa[2,3], Peter Tompa[2,3], Tom Lenaerts[1,4,5], Wim Vranken[2,3]*
[1]MLG, Département d'Informatique, Université Libre de Bruxelles, Belgium. [2]Structural Biology Brussels, Vrije Universiteit Brussel, Belgium. [3]Department of Structural Biology, VIB, Belgium. [4]AI-lab, Computer Science department, Vrije Universiteit Brussel, Belgium. [5]Interuniversity Institute of Bioinformatics in Brussels (IB²), Brussels, Belgium.
[*]wvranken@vub.ac.be

Protein function and dynamics are closely related, but accurate dynamics information is difficult to obtain. Based on a carefully assembled dataset derived from experimental data for proteins in solution, we develop DynaMine, a fast, high-quality predictor of protein backbone dynamics. DynaMine uses only sequence information as input and shows great potential in distinguishing regions of different structural organization, such as folded domains, disordered linkers, molten globules and pre-structured binding motifs. It also identifies disordered regions within proteins with an accuracy comparable to the most sophisticated existing predictors, without depending on prior disorder knowledge or structural information. DynaMine provides molecular biologists with an important new method that grasps the dynamical characteristics of any protein of interest, as demonstrated here for human p53.

INTRODUCTION

Dynamics are essential for protein function. Intrinsically disorder proteins[1] are an emblematic example of this since they function as an ensemble of conformations without a consistent three-dimensional structure. Understanding dynamics and disorder poses significant challenges, but here we introduce a new prediction method that can smartly overcome the underlying problems by the accurate prediction of protein backbone dynamics from sequence.

METHODS

By leveraging on a large set of NMR chemical shift data extracted from the BioMagResBank[2], we estimated backbone N-H S^2 order parameter values with the Random Coil Index software[3] for 218259 residues in 2015 proteins. S^2 order parameters represent how restricted the movement of an atomic bond vector is with respect to the molecular reference frame. A value of 1 means complete order (stable conformation), while 0 means fully random bond vector movement (highly dynamic). Based on these values we built DynaMine[4], a sequence-based predictor of protein backbone dynamics, based on a simple linear regression algorithm. DynaMine takes as input a protein sequence and produces a profile of per-residue predicted S^2 values (S^2_{pred}) as in Figure 1. Each S^2_{pred} is predicted based on the context provided by the 25 residues preceding and following it in the sequence. Predictive performance is evaluated by 10-fold cross-validation.

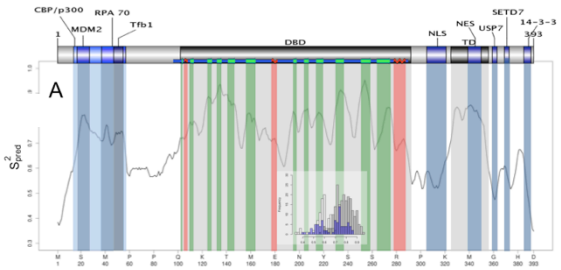


***FIGURE 1**. DynaMine S^2_{pred} values for the p53 human tumor suppressor. In the top bar, folded domains are colored in black, binding regions in shades of blue. Secondary structure is presented in the DNA Binding Domain (DBD) region (helices–red, strands–green).*

RESULTS & DISCUSSION

We demonstrate that statistical analysis of NMR data of proteins in solution can give quantitative insight into the relationship between amino acid sequence and backbone dynamics. The DynaMine backbone dynamics predictor rests on S^2 order parameters directly estimated from experimental data content (NMR chemical shifts) and produces excellent results despite the simple linear prediction methodology applied. DynaMine is very fast and gives a continuous and subtle picture of how amino acid residues behave with respect to their backbone rigidity and, by extension, to residue order and disorder (Figure 1).
We show on a set of well-studied proteins covering the full range of structural and functional properties that DynaMine has great potential in distinguishing regions of different structural organization, such as folded domains, disordered linkers, molten globules and pre-structured binding motifs (e.g. Figure 1). We also show that it can identify disordered protein regions with an accuracy comparable to the most sophisticated existing predictors, without relying on any prior disorder annotation (Figure 2). To conclude, we contend that DynaMine provides independent evidence and an unbiased picture of dynamics and structural disorder.

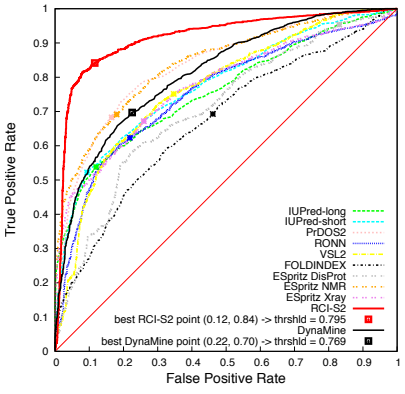


***FIGURE 2**. DynaMine performance in relation to disorder and compared with other disorder predictors on a validation dataset.*

REFERENCES

1. Tompa, P. *Trends in biochemical sciences* **27**, 527-533 (2002).
2. Ulrich, E. *et al. Nucleic Acids Res* **36**, D402-408 (2008).
3. Berjanskii, M. V. & Wishart, D. S. *Journal Of Biomolecular NMR* **40**, 31-48 (2008).
4. Cilia, E. *et al. Nature Communications* **4**:2741, (2013).

pE-DB: A Database of Structural Ensembles of Intrinsically Disordered and of Unfolded Proteins

Mihaly Varadi[1,], Simone Kosol[1], Pierre Lebrun[1], Erica Valentini[2], Martin Blackledge[3], A. Keith Dunker[4], Isabella C. Felli[5], Julie D. Forman-Kay[6], Richard W. Kriwacki[7], Roberta Pierattelli[5], Joel Sussman[8], Dmitri I. Svergun[2], Vladimir N. Uversky[9,10], Michele Vendruscolo[11], David Wishart[12], Peter E. Wright[13], Peter Tompa[1,14].*

VIB Department of Structural Biology, VUB[1]; EMBL, Hamburg, Germany[2]; CEA, CNRS, Institut de Biologie Structurale Jean-Pierre Ebel, Grenoble, France[3]; Indiana University School of Medicine; Indianapolis, IN, USA[4]; CERM, Department of Chemistry, University of Florence, Italy[5]; Molecular Structure and Function Program, Hospital for Sick Children, Toronto, Canada[6]; Department of Structural Biology, St. Jude Children's Research Hospital, Memphis, TN, USA[7]; Department of Structural Biology, Weizmann Institute of Science, Rehovot, Israel[8]; Department of Molecular Medicine and USF Health Byrd Alzheimer's Research Institute, Morsani College of Medicine, Tampa, FL, USA[9]; Institute for Biological Instrumentation, Russian Academy of Sciences, Pushchino, Russia[10]; Department of Chemistry, University of Cambridge, Cambridge, UK[11]; Departments of Biological Sciences and Computing Science, University of Alberta, Edmonton, AB, Canada[12]; Department of Integrative Structural and Computational Biology, The Scripps Research Institute, La Jolla, CA, USA[13]; Institute of Enzymology, Hungarian Academy of Sciences, Budapest[14]. [*]mvaradi@vub.ac.be

The goal of pE-DB (http://pedb.vib.be) is to serve as an openly accessible database for the deposition of structural ensembles of intrinsically disordered proteins (IDPs) and of denatured proteins based on nuclear magnetic resonance (NMR) spectroscopy, small-angle X-ray scattering (SAXS) and other data measured in solution. Due to the inherent flexibility of IDPs, solution techniques are particularly appropriate for characterizing their biophysical properties, and structural ensembles in agreement with these data provide a convenient tool for describing the underlying conformational sampling.

Intrinsically disordered proteins

Intrinsically disordered proteins (IDPs) are defined by the lack of a single, static tertiary structure under physiological conditions[1]. These proteins have multiple conformations that are separated by low free energy barriers and consequently their structures constantly fluctuate between different states giving rise to a dynamic ensemble of conformations. Disordered regions are ubiquitous in proteins involved in biological processes of DNA and RNA binding, transcription, translation, cell-cycle regulation and membrane fusion, and also often in pathologies associated with misfolding and aggregation, as observed in a variety of neurodegenerative diseases and in the pathogenesis of many other human maladies. Although a structural description of IDPs is not feasible using X-ray crystallography, other techniques, such as NMR experiments measuring chemical shifts (CSs), Residual Dipolar Couplings (RDCs), 15N R2 relaxation rates, Paramagnetic Relaxation Enhancement (PRE) distance restraints, J-couplings, 1H-15N heteronuclear Nuclear Overhauser Effects (hetNOEs) and SAXS measurements can yield meaningful information on the distribution of their shape and size, short- and long-range contacts and backbone flexibility[2,3].

Methods

Database entries consist of (i) primary experimental data with descriptions of the acquisition methods and algorithms used for the ensemble calculations, and (ii) the structural ensembles (Figure 1) consistent with these data, provided as a set of models in a Protein Data Bank (PDB) format.

Future Outlook

pE-DB is open for submissions from the community, and is intended as a forum for disseminating the structural ensembles and the methodologies used to generate them. The availability of the ensembles and the underlying data is expected to promote the development of new modeling methods and lead to a better understanding of how function arises from disordered states.

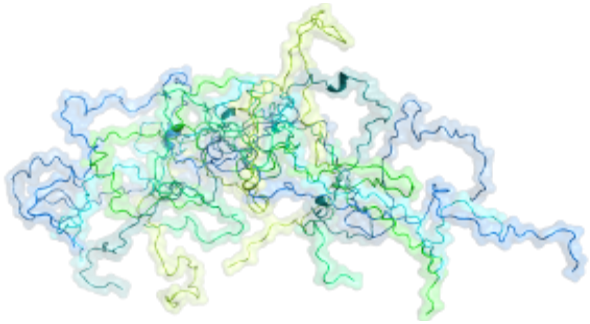


Figure 1. *Conformational ensemble of an IDP.*

References

1. Tompa, P. (2011) Unstructural biology coming of age. Curr Opin Struct Biol, **21**, 419-425.
2. Bernado, P. and Svergun, D.I. (2012) Structural analysis of intrinsically disordered proteins by small-angle X-ray scattering. Mol Biosyst, 8, 151-167.
3. Blackledge, M. *et al.* (2012) Towards a robust description of intrinsic protein disorder using nuclear magnetic resonance spectroscopy. Mol Biosyst, 8, 58-68.

THE SELECTIVITY OF THE VOLTAGE-DEPENDENT ANION CHANNEL TOWARDS INORGANIC IONS

Eva-Maria Krammer[1,*], *Fabrice Homblé & Martine Prévost*
Structure and Function of Membrane Proteins[1], *Université Libre de Bruxelles.* [*]*ekrammer@ulb.ac.be*

The voltage-dependent anion channel (VDAC) forms the major pore in the outer mitochondrial membrane. Its high conducting open state features a moderate anion selectivity that can be modulated by variations of salt concentration. We examined the translocation of small inorganic ions across mouse wild-type VDAC1 and mutated variants using different structure based theoretical methods. Our data bolster the role of the charge distribution inside the pore as the main determinants of VDAC selectivity towards inorganic anions.

INTRODUCTION

The mitochondrial respiration requires the exchange of inorganic ions and metabolites between the cytoplasm and the mitochondrial matrix. The voltage-dependent anion channel (VDAC) mediates this exchange through the outer mitochondrial membrane. The physiological significance of VDAC in the mitochondrial metabolism was suggested to be strongly correlated to its voltage-dependence[1]: at voltages close to 0 mV, the channel exists in a fully open state whereas upon higher voltages it switches to partially closed states. VDAC open state is characterized by a high conductance compatible with the magnitude of the metabolite flow through mitochondria and by a slight preference for inorganic anions over cations. In contrast, in its closed states, VDAC is no longer permeable to metabolites and shows a lower conductance for small ions with a preference for cations.

In 2008 the 3D structure (see Figure 1) of mammalian VDACs open state was solved[2] revealing a large barrel formed by 19 β-strands with an N-terminal helix folded inside. The combination of now available atomic resolution structures with computer modelling and simulation tools paves the way to unravel, at the molecular level, the fundamental principles of ion translocation through the channel.

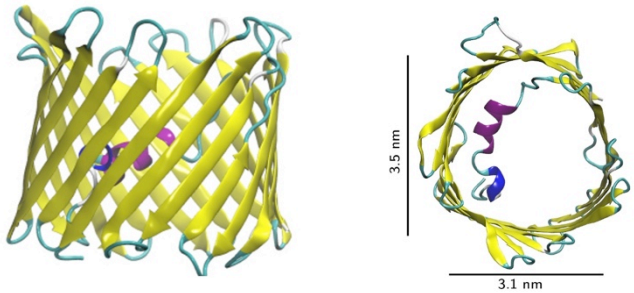


***FIGURE 1**. Side (left) and top (right) view of the mouse VDAC1 crystal structure.*

METHODS

Brownian Dynamics Simulations (BD): All BD simulations were performed on the mVDAC1[2] (PDB ID: 3EMN) crystal structure using the GCMC/BD program[3]. The reversal potential and the conductance were computed from non-equilibrium 0.5 μs BD simulations performed in 1.0 M/0.1 M (*trans*/*cis*) KCl and NaCl gradients respectively, applying a membrane potential ranging from −50 mV to 50 mV with a stepsize of 10 mV.

Molecular Dynamics Simulations (MD): All MD simulations were performed on the mVDAC1 crystal structure embedded in POPE and in the presence of different KCl concentration solutions (0.1M, 0.2M, 0.4M, 0.8M, 1.0M, and 1.2M) using the program NAMD[4]. The single mutant proteins were generated using the Mutate module in VMD[5].

Continuum Electrostatics Simulations (CE): Ion transfer free energies were calculated at 300 K for a chloride-sized anion and a potassium-sized cation through the mVDAC1 crystal structure pore at 0.1 M and 1.0 M KCl. A grid with a spacing of 1 Å was used with a size of 40×40×70 Å^3. At each point the electrostatic free energy corresponding to the transfer of an ion from solution into the protein/lipid environment was calculated using the apbs program [6].

RESULTS & DISCUSSION

The translocation of small inorganic ions through VDAC was examined using MD and BD simulations together with CE calculations of mouse VDAC isoform 1 wild-type and mutants to propose a molecular mechanism for VDAC selectivity. The analysis of the simulation trajectories indicates no distinct pathways for ion diffusion and no long-lived ion-protein interactions[7,8]. It points to a pore region comprising the N-terminal helix and the barrel band encircling it as a major controller of the ion transport across the channel. The calculated dependence of ion distribution in the wild-type channel with the salt concentration is in very good agreement with the experimental observations and can be explained by an ionic screening of the charged residues located in the pore. Altogether these results bolster the role of electrostatic features of the pore as the main determinants of VDAC selectivity towards inorganic anions.

REFERENCES

1. Colombini M, *Mol. Cell. Biochem* **256-257**, 107-115 (2004).
2. Uwjal R *et al.*, *Proc. Natl. Acad. Sci. USA* **105**, 17742-17747 (2008)
3. Im W et al. *Biophys. J.* **79**, 788-801 (2000)
4. Philipps JC *et al., J. Comput. Chem.* **26** 1781-1802 (2005).
5. Humphrey W, Dalke A & Schulten K. *J. Mol. Graph.* **14**, 33-38 (1996).
6. Holst M & Saied F, J. *Comput. Chem.* **14,** 105-113 (1993).
7. Krammer E-M, *et al.*, *Plos One* **6**, e27994 (2009)
8. Krammer E-M. *et al.*, *Biochim. Biophys. Acta* **1828**, 1284-1292 (2011).

PROTEIN FOLD RECOGNITION THROUGH HYBRID GEOMETRIC KERNEL INTEGRATION OF DIFFERENT PROTEIN FEATURES WITH COMPLEMENTARY INFORMATION

Pooya Zakeri[1,2,], Ben Jeuris[3], Raf Vanderbil[3], Yves Moreau[1,2]*
Department of Electrical Engineering – ESAT, SCD-SISTA KU Leuven, Leuven , Belgium[1], and Future Health Department, iMinds, Leuven , Belgium[2], Department of Computer Science, KU Leuven, Leuven, Belgium[3],
Pooya.Zakeri@esat.kuleven.be [*]

Tertiary structural information of proteins can provide new knowledge on their function. In addition, *understanding the three-dimensional structure* of proteins can be facilitated through the knowledge of protein folds. Various protein sequence feature-based approaches have been used in the classification of protein folds. More attention needs to be paid to find an efficient method for fusing these heterogeneous and discriminatory data sources. We propose a novel approach to integrate kernel matrices though taking Log-Euclidean mean instead of convex linear combination. We integrate 26 different representative models of protein domains based on physicochemical properties and primary and secondary structural information of protein sequences, as well as information extracted from local sequence alignment and position specific scoring matrices. Moreover, we expand our computational model by combining the available knowledge on functions of protein domains through the hybridization model.

INTRODUCTION

Early and late integrations are common approaches to integrate various proteins fold data sources. In addition, the heterogeneous biological data sources can be fused intelligently using kernel-based data fusion. The main intuition behind the kernel integration approaches is to first construct the same representation for all data such as kernel matrices and then integrate these representations systematically. The standard approach for combining kernel matrices is to take the weighted arithmetic average. This type of averaging is often sensitive to deal with complementary and noisy kernels, which is typical situation in biological applications. This motivates us to think about the geometric mean between symmetric positive definite kernel (SDP) matrices which are not necessarily a linear combination of SPD matrices and relative to Euclidean distance on a convex cone whose interior contains all SPD matrices.

METHODS

However, computing the geometric mean for a general number of SDP matrices is hard and computationally expensive, which is why we also discuss the Log-Euclidean mean [3]. It can be considered as a consensus between the arithmetic and geometric mean. The Log-Euclidean mean of *n* SPD matrices can be obtained explicitly as [3]:

$$K_{LE}(K_1,\ldots,K_n)=\exp\left(\frac{1}{n}\left(\sum_{i=1}^{n}\log(K_i)\right)\right)$$

We evaluate our method for classification on the SCOP PDB-40D benchmark dataset [1] which consists of 27 SCOP fold classes. Gaussian RBF kernel function is employed for 26 different protein features. We fuse 26 RBF kernels derived from each view on protein domains through taking Log-Euclidean mean of them. Then the classification is performed using a one-versus-rest (OVR) support vector machine (SVM) (LogFold). To combine the available functional domain composition (FunD), we consider the FunD composition of protein sequences using the InterPro database [2]. Then the prediction will be perform by combing FunD kernel [4] and fused kernel produced by Geofold model (FunGeoFold). The general architecture of the proposed approaches is illustrated in Figure 1.

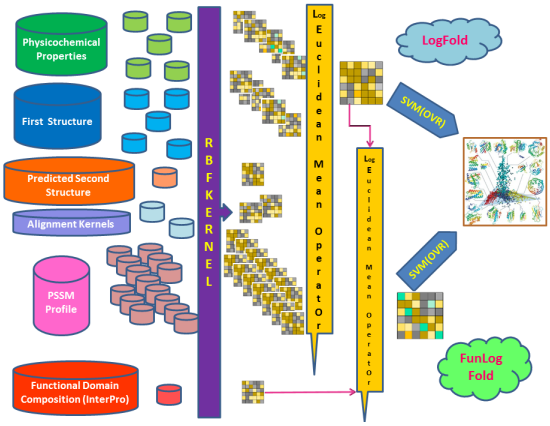


FIGURE 1. The architecture of our integration model.

RESULTS & DISCUSSION

The total accuracy on the protein fold recognition test set obtained by our methods is about 81.7%. Moreover, The protein fold recognition accuracy is improved to 90% by integrating the functional domain composition through our proposed hybridization model. These results indicate that the Log-Euclidean mean of base kernels can effectively improve the accuracy of the state-of-the-art kernel fusion model for protein fold recognition. Furthermore, the 60.6% performance of uniformly weighted linear integration of 26 base kernels indicates the limitation of convex linear combinations in dealing with integrating of different protein features which carry complementary information.

REFERENCES

1. Ding, C. H. and Dubchak I. *Bioinformatics* **17**, 349-358 (2001).
2. Hunter S *et al*. *Nucleic Acids Res* **40**, 306-312 (2012).
3. Arsigny, V. *et al SIAM J Matrix Anal A*,**29(1)**, 328-347 (2007).
4. Zakeri P. *et al* J *Theor Biol*, **269(1)**, 208-216 (2011).

PROTEIN THERMAL STABILITY PREDICTION BY STATISTICAL POTENTIAL

Fabrizio Pucci[1,*], *Malik Dhanani*[1], *Yves Dehouck*[1], *Marianne Rooman*[1]

Department of BioModeling, BioInformatics & BioProcesses[1], *Roosevelt Avenue 50, 1050 Brussels, Belgium.*

**fapucci@ulb.ac.be*

A lot of in silico methods have been developed to analyze the thermodynamic stability of proteins. In contrast, no method exists to directly predict the best descriptor of thermal stability, that is, the melting temperature T_m, on the basis of the protein sequence and structure. Here we derive and test a new melting temperature prediction method that is applicable to families of homologous proteins. It is based on a series of (melting)temperature-dependent statistical potentials derived from ensembles of mesostable and thermostable proteins. It performs significantly better than methods that predict T_m from thermodynamic stability: in the test set of 45 proteins distributed among 11 families, the standard deviation computed in cross-validation between the predicted and the experimentally measured melting temperature is about 13 °C.

INTRODUCTION

Even if significant progress has been made in the last decade, the understanding of protein thermal stability remains one of the key questions of protein science. Indeed it is still highly non-trivial to have precise predictions about thermal stability. The results are in general family dependent and sometimes even contradictory. A better comprehension of this issue and improved predictions would have many applications, for example in the optimization of protein-based bioprocesses and in the design of new drugs.

Here we develop a tool for the prediction of the best descriptor of the thermal stability, namely the melting temperature T_m, corresponding to the temperature at which the protein denatures. We limited our predictions to eleven families of homologous proteins (in total 45 proteins), which represent all the protein families containing at least 3 crystal structures with experimentally determined T_m.

METHODS

To derive the prediction methods for the melting temperature we use as a tool the statistical potentials already described in literature and applied to the computation of changes in folding free energy upon mutation[1,2]. The basic idea behind the construction of such potentials is to compute the relative frequencies of sequence and structure motifs in a given protein dataset and convert them, trough the Boltzmann law, into folding free energy.

To analyze the thermal stability, the standard construction[1] has to be modified so as to consider that the amino acid interactions are temperature dependent : some of them could be more stabilizing than others at high temperature and vice versa.

To take into account such dependence we created different datasets containing only proteins with melting temperature in a certain range : a first set of mesostable proteins (T_m < 65 °C) and a second set of thermostable proteins. A third set, used as reference, contains all the proteins independently of their T_m. From these sets, three different potential are derived, the mesostable ΔG^m, the thermostable ΔG^t and the reference potential ΔG^r.

The eleven families of homologous protein we consider are α-amylase, lysozyme, myoglobin, β-lactamase, α-lactalbumin, acylphosphatase, adenylate kinase, cell 12A endoglucanase, cold shock protein, cytochrome and ribonuclease.

We computed the folding free energy of the 45 collected protein using the temperature dependent potential ΔG^t - ΔG^m and the standard one ΔG^r . Finally, from these folding free energy values, the melting temperature of the protein has been predicted trough a simple linear extrapolation.

RESULTS & DISCUSSION

We computed for the eleven families of homologous proteins, the standard deviation σ between the experimentally measured melting temperature and the predicted one. Using the temperature dependent potentials, we found a value of σ of about 13 °C. This value is much better than the one obtained from the reference potential ΔG^r -thus from thermodynamic stability predictions-, which yields a σ of about 18 °C. Moreover, if, on the total of 45 proteins analyzed, we remove the six proteins that are predicted worst, the standard deviation reduce to 8°C. This is quite a good accuracy compared to the experimental and computational errors.

Some optimization of the method are still possible. However, the main problem that affects the results is the lack of data; the number of proteins with known structure and melting temperature is indeed much too small.

REFERENCES

1. Y. Dehouck, D. Gilis, M. Rooman (2006), A new generation of statistical potentials for protein, *Biophys. J.* **90**, 4010–4017.
2. Y. Dehouck, Jean Marc Kwasigroch, D. Gilis, M. Rooman (2011), PopMusic 2.1 : a web server for the estimation of the protein stability changes upon mutation and sequence optimality. *BMC Bioinformatics.* **12**, 151
3. B. Folch, Y. Dehouck, M. Rooman (2010), Thermo- and mesostabilizing protein interactions identified by temperature-dependent statistical potentials, *Biophys. J.* **98**, 667-77.

COMPUTATIONAL ANALYSIS OF ANTIGEN-ANTIBODY COMPARED TO OTHER PROTEIN-PROTEIN INTERACTIONS

*Giorgos Dalkas, Fabian Teheux, Jean Marc Kwasigroch, Marianne Rooman**
Department of BioModeling, BioInformatics & BioProcesses, Université Libre de Bruxelles, 50 Roosevelt Ave, 1050 Brussels, Belgium. *mrooman@ulb.ac.be*

The identification and prediction of immunogenic regions on the antigen surface that can stimulate immune response is one of the major challenges for the design and development of new vaccines. Several computational methods have been developed in recent years for predicting potential B-cell epitopes, but their predictive performance remains unsatisfactory. A key step that could contribute to improve B-cell epitope prediction tools is a better understanding of the interactions that stabilize the antigen-antibody complexes. For this purpose, we identify the residue-residue contacts that occur across the antigen-antibody interfaces, and compare them with those occurring in other protein-protein interfaces. Several interactions, such as cation-π, amino-π and π–π interactions, were found to be much more frequent at antigen-antibody interfaces.

INTRODUCTION

Humans are constantly exposed to a wide variety of pathogens and infectious agents; the role of the immune system is to protect them against these infections. A crucial part of the humoral immune responses is the B-cells, which secrete antibodies against antigens. Antibodies bind to antigens at sites called B-cell epitopes. The identification and characterization of epitope regions on the antigen surface, which are capable of inducing an efficient immune response, is one of the key steps for the development of new vaccines[1]. Most of the existing prediction tools are based on studies indicating that protein-protein interfaces differ in their features (*e.g.* amino acid composition) from the remaining protein surface[2]. It is also generally assumed that antigen-antibody interfaces are different from other protein-protein interfaces[3]. To further investigate this, we performed a systematic analysis of the differences in amino acid composition and secondary structure between antigens, antibody and other protein surfaces, as well as between epitopes, paratopes and other protein-protein interfaces. The observed variations were correlated with differences in residue-residue contacts across the interfaces.

METHODS

To study the antigenic properties of proteins, a dataset of experimentally determined antigen-antibody structures was obtained from the IEDB-3D, the 3D structural component of the Immune Epitope Database. This dataset was filtered using several criteria, such as resolution and sequence similarity. The antibody–antigen complexes were separated into two groups, taking into account the structural epitope segmentation. Epitopes consisting of a continuous stretch of residues that may include a gap of up to three non-epitope residues were considered as linear epitopes. The other epitopes, which contain distant amino acids that are brought together by the folding of the polypeptide, were considered as conformational epitopes. Finally, a dataset of protein–protein heterodimers was used for comparison and reference.

The residues involved in the interfaces where identified on the basis of their change in solvent accessibility upon binding. The amino acid composition of the interfaces and the types of interactions that link residues across the interfaces were investigated. These interactions include salt bridges, disulfide bonds, H-bonds, hydrophobic packing, as well as π–π, cation-π and amino-π interactions.

RESULTS & DISCUSSION

The amino acid composition of the conformational B-cell epitopes was found to be enriched in aromatic amino acids and in residues carrying a partial or net charge, and depleted of aliphatic residues, in comparison with the rest of the antigen surface. Compared to conformational epitopes, linear epitopes contain less charged and more aliphatic residues, as well as more Pro and Gly. Other protein-protein interfaces show an even larger amount of hydrophobic residues. The observed amino acid composition in the different types of interfaces can be put in parallel with the interactions that stabilize them. Indeed, the analysis of the antigen-antibody and protein-protein binding interactions revealed the significant preference of cation-π, amino-π and π–π interactions in antigen-antibody complexes, whereas other protein-protein complexes are more stabilized through hydrophobic packing.

These findings will be exploited to improve the performance of B-cell epitope prediction methods.

REFERENCES

1. Irving MB *et al*. *Curr Opin Chem Biol* **5**, 314–324 (2001).
2. Jones S & Thornton JM. *J Mol Biol* **272**, 133–143 (1997).
3. Lo Conte L *et al*. *J Mol Biol* **285**, 2177–2198 (1999).

PAIRWISE KERNEL METHODS FOR PREDICTING MOLECULAR INTERACTIONS

Michiel Stock[1,*], *Tapio Pahikkala*[2], *Antti Airola*[2], *Bernard De Baets*[1] & *Willem Waegeman*[1].
KERMIT[1], *Department of Mathematical Modelling, Statistics and Bioinformatics, Ghent University;*
Department of Information Technology and the Turku Centre for Computer Science[2], *University of Turku.*
[*]*Michiel.Stock@UGent.be*

An important part of bioinformatics is devoted to predicting interactions between the molecular components in a cell. For example in drug design, one is interested to know whether a ligand will bind to a certain protein. To understand epigenetic regulation, the hybridisation between micro and messenger RNA is of interest. As the physicochemical mechanisms of these processes are often insufficiently understood, statistical models are more suited to make reliable predictions. We present a very general framework based on constructing a pairwise Kronecker product kernel, which allows us to predict properties of a pair of objects. Our methods can deal with the typical complex data types that are often encountered in bioinformatics and are efficient enough to handle realistic problems.

INTRODUCTION

In bioinformatics a shift of focus has occurred from studying individual biomolecules, such as proteins, nucleic acid sequences or metabolites, to systems of interacting components. Being able to predict the interaction between these molecules *in silico* is an important first step towards a systems biology approach.

METHODS

Many machine learning methods exist for learning relations between objects. Ours are based on a joint feature representation of pairs of objects using the Kronecker product pairwise kernel (KPPK). This KPPK is constructed by combining individual object kernels using a Kronecker product, as shown in Figure 1. This frameork is popular in bioinformatics since it was first introduced to predict protein-protein interactions[1]. Waegeman et al.[2] have given this paradigm a more theoretical foundation.

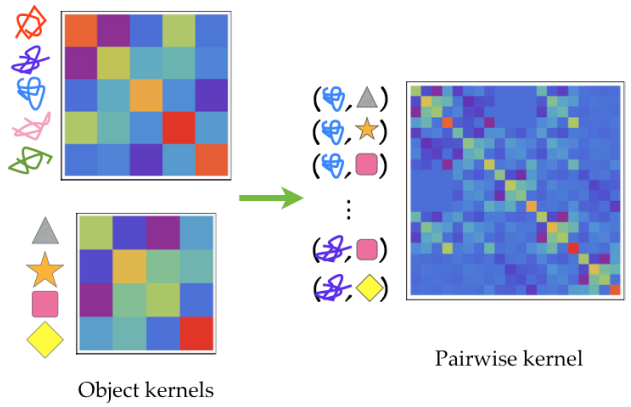


***FIGURE 1**. Using a kernel comparing five proteins and a kernel describing four ligands, one can use the Kronecker product to construct a protein-ligand kernel.*

Kernel methods have proven to be a useful instrument in the bioinformatician's toolkit[3]. One reason is because they provide a natural way to represent complex structures such as sequences, strings, graphs or trees, which are omnipresent in biology. A second reason is that a relevant (pairwise) kernel matrix representation can be plugged into any of the available supervised or unsupervised learning algorithms.

RESULTS & DISCUSSION

We have applied our method in a variety of bioinformatics problems: predicting enzyme function, modelling protein-ligand interactions and miRNA-mRNA interactions. Often, the performance is comparable to the state-of-the-art. The KPPK is a very natural way to approach these problems.
Building a statistical model for predicting properties of two objects is significantly more complex than when considering only one object. When tuning the hyperparameters one should ask if the goal is to predict for new combinations of known objects or if the goal is to predict for completely unseen objects, or even a setting in between. These distinctions are very important from an application point of view, but still lack a theoretical understanding[4]. Also, the most relevant way of performing model evaluation, for example, by means of the area under the ROC curve, is still under investigation. Currently, we are performing a series of computer simulations in order to gain more insight in these both fundamental and practical questions. These indicate that controlling the dependence between the objects is vital.
A common critique on these methods is that it is very demanding to compute calculate and process the KPPK matrix. Though, if one uses methods with a quadratic loss function, such kernel ridge regression, it can be shown that the most computer intensive operation needed is the eigenvalue decomposition of the individual object kernel matrices. The calculation is thus $O(n^3+m^3)$. This makes the KPPK accessible for realistic bioinformatics applications with thousands of objects and millions of interactions. In most cases one does not lose any accuracy by using a simple kernel ridge regression, compared to more complex methods such as support vector machines.

REFERENCES

1. Vert, J-P *et al. BMC Bioinformatics*, **8** (Suppl 10):S8, (2007).
2. Waegeman, W *et al. IEEE Transactions on Fuzzy Systems.* **99**, 1 (2012).
3. Schölkopf, B *et al.* Kernel Methods in Computational Biology. The MIT Press (2004).
4. Park, Y and Marcotte, E. *Nature Methods,* **9**, 1134-1136, (2012).

CAPRI: THE DIVERSE CHALLENGES OF COMPUTATIONAL PROTEIN-PROTEIN DOCKING

Marc F. Lensink
Molecular Systems Biology, Interdisciplinary Research Institute, CNRS USR3078, University Lille Nord of France.
marc.lensink@iri.univ-lille1.fr

Protein-protein interaction (PPI) lies at the core of cellular functioning. Computational protein docking is the process of obtaining the three-dimensional coordinates of a macromolecular assembly. The CAPRI experiment is a community-wide collaboration aimed at the improvement of computational protein docking. In the 10 years of its existence it has catalyzed the development of docking algorithms and methodologies and led to a fundamental improvement of our understanding of PPI.

BACKGROUND

Protein-protein interaction (PPI) is omnipresent in life and plays a central role in all biological processes. These processes result from the physical interaction of two or more protein molecules, forming the macromolecular assemblies that perform many of the cellular functions. Many large-scale studies focusing on PPI have emerged in recent years, displaying cobweb-like networks to represent the protein components interacting with each other. Owing to the sheer massiveness of the underlying data, such representations capture a wealth of useful information, but in spite of this usefulness, the representation itself remains abstract and incomplete; no information as to time, place or specificity is included.

PROTEIN DOCKING

Such detailed information is indispensable for the guidance of mutagenesis studies or the design of inhibitor molecules. The actual formation of the three-dimensional protein complex is the topic of computational protein docking. Present-day realistic scenarios involve the use of unbound template structures, requiring a step of homology modeling prior to docking.

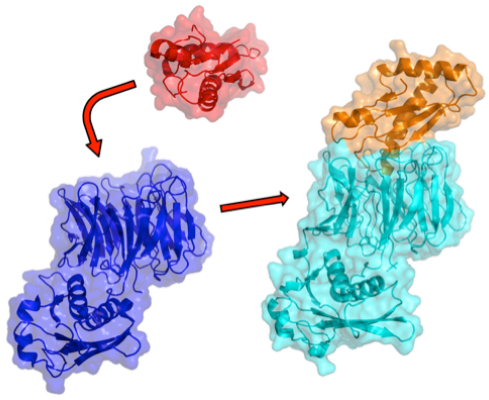


***FIGURE 1**. Computational protein docking produces the three-dimensional coordinates of a protein complex (orange and cyan for bound ligand and receptor entities), starting from the individual structures as they are found in solution (red and blue for unbound ligand and receptor entities).*

CAPRI

The CAPRI (Critical Assessment of PRedicted Interactions) experiment was initiated some 10 years ago[1]. Modeled after CASP (Critical Assessment of protein Structure Prediction), participants in CAPRI are asked to predict the three-dimensional structure of a protein complex; these predictions are then assessed in a double-blind procedure against an unpublished and confidential X-ray target structure. Participants in CAPRI meet at a regular basis at Evaluation Meetings; six of these meetings have been held so far, the most recent one in Utrecht in April, 2013. The results of the assessment and contributions by a selection of predictor groups are published periodically in a CAPRI-dedicated issue of the journal Proteins[2,3].

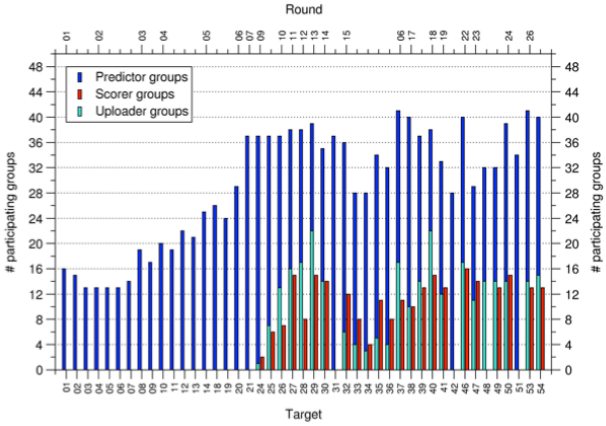


***FIGURE 2**. Number of participating research teams in CAPRI since its inception. Predictor groups (blue bars) participate in the docking experiment, scorer groups (red bars) participate in the scoring experiment, selecting models made available by uploader groups (cyan bars), who form a subset of the predictor groups.*

A DYNAMIC AND DIVERSE EXPERIMENT

The targets in CAPRI follow the demand of the experimentalists and as such represent a wide variety of biological processes. The latest CAPRI assessment featured the most diverse set of complexes ever encountered. The experiment now includes the prediction of multi-component assemblies, protein-nucleic acid and protein-polysaccharide binding[2,3], but also of binding affinities[4] and the positions of interfacial water molecules[5]. The CAPRI project gives a fair assessment of the performance of present-day protein docking methods for the more difficult targets and as such it provides an upper limit of what can be expected from docking algorithms.

REFERENCES

1. Janin J *et al*. *Proteins* **52**, 2-9 (2003).
2. Lensink MF & Wodak SJ., *Proteins* **78**, 3073-3084 (2010).
3. Lensink MF & Wodak SJ, *Proteins*, in press (2013).
4. Moretti R et al., *Proteins*, in press (2013).
5. Lensink MF *et al*, *Proteins*, in press (2013).

IDENTIFYING DEGRONS: HOW ARE PROTEINS TARGETED FOR DEGRADATION BY THE UBIQUITIN PROTEASOME SYSTEM?

Pallab Bhowmick[1], Mainak Guharoy[1,] & Peter Tompa[1,2*].*
VIB Dept. of Structural Biology, Vrije Universiteit Brussel[1] and Institute of Enzymology, Budapest[2].
[*]PT: ptompa@vub.ac.be; [*]MG: mainak.guharoy@vib-vub.be

Cellular regulation requires the targeted degradation of proteins when they are no longer needed. The process of recognizing target proteins is highly specific, and involves E3 ubiquitin ligase enzymes that label target proteins with poly-ubiquitin chains (the signal for degradation). Approximately 600 human E3s interface with the proteome in a set of highly specific reactions. In this work, we describe the occurrence of peptide motifs within proteins that serve as the target site for E3 recognition ('degrons'). We describe the sequence and structural features of degrons and outline a computational protocol for identifying degron signatures in candidate proteins.

INTRODUCTION

The ubiquitin–proteasome system (UPS) is an evolutionarily conserved and central component of cellular regulation. The system works as an enzymatic cascade (built as a pyramid) with two E1 (ubiquitin activating), few dozen E2 (ubiquitin conjugating) and several hundred E3 (ubiquitin ligase) enzymes[1]. The E3s are directly responsible for specific recognition of target proteins and then labeling them with poly-Ubiquitin (Ub) chains. Understanding the molecular basis of specific E3-substrate recognition is critical for a basic knowledge into this key cellular pathway, but more importantly for insights into disease conditions due to impaired E3 functioning. Of particular interest are impaired recognition and degradation of cellular growth factors that lead to tumor progression.

We demonstrate in this work that E3s recognize highly specific short linear motifs within target proteins, often only exposed on the surface after post-translational modifications (PTMs) (the 'turn-off' switch signal). Further, we describe sequence and structural features of these degradation motifs ('degrons'); the features are then used to train a predictor for identifying E3-binding degrons within protein sequences. Lastly, we plan to use the predictor to scan the whole human proteome for putative degron signatures. Some of the candidates will be taken to an experimental validation step.

METHODS

First, we obtained from the literature a set of 8 human E3 ligases (eg, MDM2, CBL), and their corresponding target substrates, for which degrons have been experimentally validated. Then we studied the sequence and structural features of these degrons and integrated these features into a general computational protocol that would allow us to derive a degron from a set of input (substrate) protein sequences. Our methodology was benchmarked for accuracy by its ability to recover the 8 experimentally validated motifs. Briefly, the method comprises the following steps:

- Substrate sequences were provided as input to the motif predictor MEME[2], and an initial library of 50 sequence motifs were generated for each E3.
- Multiple parameters were then computed for each candidate motif: evolutionary conservation, disorder scores, probability of forming disordered binding site, surface accessibility, and occurrence of PTMs.
- These features were tested (individually, and in combination) for their ability to enrich the true (known) degron motif from the initial MEME pool.
- Finally, a scoring system combining the individual scores for each parameter was devised and trained into a support vector machine (SVM).

RESULTS & DISCUSSION

Specific binding of E3s to degradation substrates proceeds via the recognition of degrons on the substrate proteins. These degrons have certain characteristic properties: they often occur within unstructured protein regions, and show significant overlap with disordered binding sites (mdm2/p53). A large majority of degrons are structurally buried, but PTMs within (or close to) the degron exposes it, ready for E3 binding. Furthermore, degrons exhibited a high degree of sequence conservation. By combining a motif detection algorithm (MEME[2]) with these features, we were able to successfully identify the true (experimentally validated) motif for the benchmark set of 8 E3s. Using this methodology, the known motifs were consistently present in the top 5 hits for each E3.

How prevalent are degrons? To answer this question, we took a list of 560 human E3s from our previous work[3] and collected the high-confidence interaction partners (from the STRING database[4]) for each E3. Taking each interaction partner, we mined the ubiquitination literature for evidence that these proteins are indeed degraded via the UPS pathway. We created a final list of 38 E3s with at least two interaction partners that are also experimentally validated substrates. These substrate sequences were then input to MEME and the initial motif pool was generated. Our predictor was then run to enrich for 'true' motifs from this pool based on the characteristic features of experimentally validated degrons. Finally, we obtained for each E3, a set of candidate motifs (top5 hits) that have high probability of being *bona-fide* degrons. This aspect is currently under progress and we plan to take some of the candidate degrons for experimental validation.

REFERENCES

1. Hershko A & Ciechanover A *Annu Rev Biochem* **67**, 425-479 (1998).
2. Bailey TL *et al*. *Nucleic Acids Res* **34**, W369-W373 (2006).
3. Bhowmick P *et al*. *PLoS One* **8**, e65443 (2013).
4. Szklarczyk D *et al*. *Nucleic Acids Res* **39**, D561-D568 (2013).

SUPERCLUSTEROID 2: THE EASY-TO-USE TOOL TO ANALYZE YOUR PROTEIN-PROTEIN INTERACTION DATA

Charalampos Moschopoulos[1,2,*], *Konstantinos Theofilatos*[3], *Spiros Likothanassis*[3], *Yves Moreau*[1,2].
Depts of Electrical Engineering (ESAT)-STADIUS[1] *and iMinds Future Health*[2]*, KU Leuven; Dept of Computer Engineering and Informatics, University of Patras, Greece*[3]. [*]*Charalampos.moschopoulos@esat.kuleuven.be*

Superclusteroid is a web tool designed to process protein-protein interaction (PPI) data and can be used for protein complex detection or for defining protein function. In its new version new modules have been added in order to increase its usefulness in Bioinformatic community. More specifically, a graph comparison module, a new visualisation plugin and a connection to the Gene Set Enrichment Analysis (GSEA) tool have been added. Moreover, recorded protein complexes of human organism hosted in the CORUM database has been added as benchmarks for evaluating the derived protein complexes candidates. Each of these services, like previous implemented modules, can be used in a serial manner or individually, according to user needs. Superclusteroid is available at: http://superclusteroid.ceid.upatras.gr/

INTRODUCTION

Superclusteroid[1] a) uploads and manipulates input of PPI data, (b) performs clustering on PPI data using 4 different algorithms, (c) visualizes PPI networks and clustering results and (d) predicts protein function. All Superclusteroid services are supported by web services in order to be part of workflows with other web Bioinformatic tools.

In this update, new modules have been added which enables Superclusteroid to be applied on PPI datasets of different organisms, while simultaneously offers more features and information to its users.

METHODS

The new modules that have been added to Supperclusteroid are the following:

- **Graph comparison**. The user can compare two different PPI datasets and retrieve information about their similarity. Text files, which include the intersection, the union or the difference between these datasets, are generated. Also, statistics like Jacard coefficient, clustering coefficient or P-values are calculated. Additionally, degree distribution and basic graph properties of each PPI network is provided.
- **Visualisation**. In order to visualize the derived clusters of Superclusteroid analysis, we used the Cytoscape web module[2], which is a Flash component with a Javascript API. This module is extremely easy to use and efficient on visualising graphs onto web pages.
- **GSEA module**. GSEA[3] tool enables Superclusteroid users to test the statistical significance of the derived protein complex candidates concerning a specific phenotypic class. GSEA is hosted as a Java component on Superclusteroid web pages.
- **New evaluation procedure**. After the appliance of a clustering algorithm, the users can test the derived protein complex candidates against the recorded protein complexes of human organism hosted by CORUM database[4]. Users can also upload their desired evaluation dataset in order to test their results.

RESULTS & DISCUSSION

A typical workflow of the Superclusteroid tool includes the following steps: the user uploads his PPI dataset in one of the 4 available formats. Then the user can perform clustering by using 4 well known algorithms, or compare his dataset against other PPI datasets of the same or different organism.

If the clustering module is chosen, then the user has the options to visualise the whole network (or each derived graph individually), to evaluate the results against recorded protein complexes of yeast or human organism or check a selected derived cluster with GSEA tool. In the visualisation module, the user can check each the function of a protein or apply basic graph functions on the PPI network. The functional categories used are those provided in the FunCat database[5].

Also, clustering statistics can be calculated for every selected cluster. Furthermore, special care was taken to make Superclusteroid interface as user friendly as possible. In every page the user can use demo data and there are help pages in every step. Finally, every step results can be downloaded in various file formats.
The tool is implemented in UNIX environment and is written in Perl. In addition to the website, web services utilizing the SOAP protocol are also available in order to design workflows and integrate them with other available resources.

To conclude, Superclusteroid is a useful tool for researchers in the life science field designed to cater all aspects of PPI data such as clustering, visualization and function prediction. After this update, Superclusteroid offers a complete environment for analyzing PPI datasets and can be easily used in a Bioinformatic analysis pipeline. Further information about the algorithms, the web services and the rest modules are offered online in the help page.

REFERENCES

1. Ropodi A *et al. EMBnet.journal* **17,** 10-15, (2011).
2. Lopes CT *et al. Bioinformatics* **26**, 2347-8, (2010).
3. Subramanian A *et al. Proc Natl Acad Sci U S A* **102**, 15545-50, (2005).
4. Ruepp A *et al. Nucleic Acids Res* **38**, D497-501, (2010).
5. Ruepp A et al. *Nucleic Acids Res* **32**, 5539-5545, (2004).

IMPROVING THE DETECTION OF BIOLOGICALLY MEANINGFUL CLUSTERS IN PROTEIN INTERACTION NETWORKS THROUGH INTEGRATED FUNCTIONAL ANALYSIS

Haiying Wang[1], *Huiru Zheng*[1], *Francisco Azuaje*[2,*]

School of Computing and Mathematics, University of Ulster[1]*; NorLux Neuro-Oncology Laboratory, Public Research Centre for Health (CRP-Santé), Luxembourg, Luxembourg, francisco.azuaje@crp-sante.lu*[2]

Despite significant advances, the detection of informative clusters in protein interaction networks faces important challenges, including the need to aid researchers in the prioritization of hundreds or even thousands of results. To address this need, we developed a method, SimTrek, for the selection of network clusters based on the analysis of their functional homogeneity. Our method substantially reduces the space of potentially spurious clusters. In this study, SimTrek improved protein complex identification regardless of network clustering approach.

INTRODUCTION

The recognized significance of the extraction of functional modules from biological networks has triggered efforts in across research communities. Various network clustering approaches have been developed and implemented, including "Clustering with Overlapping Neighborhood Expansion" (ClusterONE), "Module Identification in Networks" (MINE), and "Markov Clustering" (MCL). Notwithstanding significant advances, the identification of biologically meaningful clusters extracted from networks faces important challenges, including the need to aid researchers in the prioritization of hundreds or even thousands of clusters. To address this need, we developed a method, SimTrek[1], which allows researchers to rank and select potentially biologically meaningful network clusters based on the analysis of their functional homogeneity, which provides objective quality criteria that are independent of the information required for obtaining the clusters.

METHODS

Based on mapping each protein interaction onto functional similarity networks inferred from Gene Ontology (GO) information, SimTrek estimates homogeneity score, Hom, associated with each cluster, which is defined as the mean of the functional similarity values observed among the interacting proteins in the cluster. The functional similarity between a pair of proteins is estimated with information encoded in the GO hierarchies: Molecular Function (MF), Biological Process (BP) and Cellular Component (CC). We assigned a statistical significance, Q, value to each Hom *s*core based on a random permutation test (Figure 1(a)). This means that relatively strongly and weakly homogeneous clusters can be distinguished by examining the functional similarity between pairs of interacting proteins. Thus, biologically relevant clusters are expected to have larger Hom scores and low Q values.

We developed SimTrek as a platform-independent Java application, which is released under the terms of the GNU Lesser General Public License. The code and Cytoscape plugin are available at: http://apps.cytoscape.org/ and http://rosalind.scm.ulster.ac.uk/Index.html.

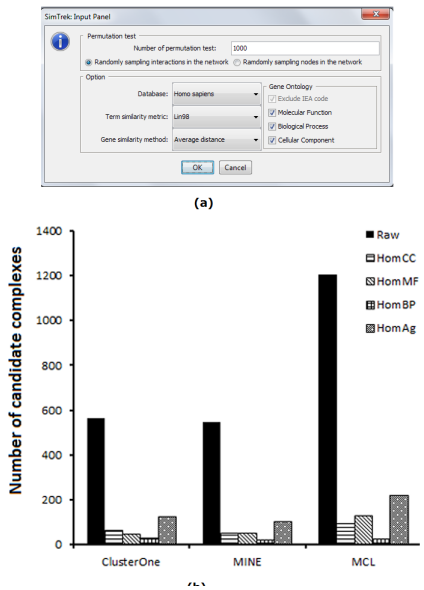


***FIGURE 1**. Overview of SimTrek. (a) Screenshot of user-interface under Cytoscape; and (b) Analysis of candidate protein complexes with SimTrek.*

RESULTS & DISCUSSION

We applied our prioritization algorithm to a global human PPI network consisting of 45901 edges (interactions) and 10653 nodes (proteins). We carried out analyses on three network clustering algorithms for candidate protein complex detection: ClusterONE, MINE, and MCL. After applying SimTrek to the raw clustering results (Hom $\geq$ 0.5 and Q $\leq$ 0.05), the space of potentially irrelevant predictions is considerably reduced in all experiments (Figure 1(b)). To independently evaluate the protein complex prediction capability of raw and SimTrek-prioritized results, Jaccard and precision-recall coefficients were estimated in relation to the reference dataset derived from the CORUM repository[2]. In the vast majority of the scenarios considered, SimTrek improved the overall capacity to detect true protein complexes by focusing on clusters reporting Hom scores with Q $\leq$ 0.05.

REFERENCES

1. Wang *et al*. *Bioinformatics*, **26,** 2643-2644 (2010).
2. Ruepp *et al.* Nucleic Acids Res., **36**, D646–D650 (2008).

INTEGRATIVE ANALYSIS OF REGULATORY TRACKS FOR THE IDENTIFICATION OF DIRECT TF-TARGET INTERACTIONS

Hana Imrichová[1*], *Gert Hulselmans*[1], *Delphine Potier*[1], *Stein Aerts*[1].
Laboratory of Computational Biology, Center for Human Genetics, University of Leuven, 3000, Leuven, Belgium[1].
[*]*Hana.Imrichova@med.kuleuven.be*

Today, more and more regulatory data is being generated using next-generation-sequencing (NGS), studying genome control in both cancer and normal samples. These datasets include genome-wide DNase-Seq (DHS), FAIRE-Seq, transcription factor (TF) binding and histone modifications by ChIP-Seq across many different cell types and conditions. A key challenge in the field of regulatory genomics remains how to optimally use these publicly available datasets for studying transcriptional regulation. We aim to tackle this challenge by an integrative approach, combining (i) genome-wide regulatory data, (ii) cis-regulatory sequence analysis, and (iii) gene expression data. For this purpose we developed a method called *i-cisTarget2.0*. Here we present a preliminary version of this tool, which allows to score gene modules and to identify upstream regulators for a given set of genes or loci, as well as cis-regulatory elements where these regulators bind.

INTRODUCTION

Thanks to the massive generation of publicly available regulatory data reflecting the activity of the non-coding genome (ENCODE[1] and other databases), it becomes possible to use a large collection of regulatory tracks to get regulatory information across many different cell lines and conditions. To handle this information in the best possible way we are developing a tool called *i-cisTarget2.0,* which integrates TF motifs with regulatory tracks. The method takes either co-regulated gene loci or co-expressed gene sets as an input, and aims to identify candidate upstream regulators alongside with direct transcriptional targets and the actual cis-regulatory modules (CRM).

METHODS

We have developed a preliminary version of a tool, called *i-cisTarget2.0.* Our analyses rely on candidate regulatory regions that we defined using publicly available regulatory data: DHS from ENCODE[1], General Binding Preference models[2], CpG islands, proximal promoters, conserved non-coding sequences, ultraconserved elements, regulatory elements from OregAnno[3], VistaEnhancers[4] and predicted cis-regulatory modules[5]. Our complete procedure of creating candidate regulatory regions yielded 1,223,024 regions (representing ~35% of the genome) with average size 818 bp. We scored and ranked these regulatory regions with all 1,849 regulatory tracks (DHS, FAIRE, ChIP-Seq) as well as with a collection of 6,661 TF motifs.

Our tool requires as an input a set of co-expressed genes or loci and provides a list of the most enriched/correlated features (motifs, regulatory tracks) as an output. In this way the input set is divided into true positive sets (sets of cis-regulatory elements where upstream regulators bind to control the expression of gene modules) and false positives. We implemented a Galaxy plugin to make *i-cisTarget2.0* available.

RESULTS & DISCUSSION

We first selected 1,223,024 candidate regulatory regions in the human genome (see Methods). We scored and ranked these regions offline with 1,849 regulatory tracks and 6,661 TF motifs, generating a large database with rankings. This database is used by a fast rank-based enrichment method to identify the most enriched features for a given set of loci.

We validated *i-cisTarget2.0* using MSigDB[6] TF perturbation gene signatures as an input to verify whether the motif and the ChIP-Seq data for the expected TF were detected. As result, *i-cisTarget2.0* detected the correct motif and/or the ChIP-Seq track for 29 from 42 TFs.

Next, we applied our method on in-house data (ChIP-Seq against p53, H3K27Ac ChIP-Seq, RNA-Seq and FAIRE, all before and after p53 activation in the MCF7 cell line). For example, using as input a set of 801 up-regulated genes after p53 activation, we could identify the corresponding p53 ChIP-Seq, active histone marks, DHS, FAIRE (all in the MCF7) and p53 motifs as top-ranked features among all thousands of tested features. This lead to the prediction of 1,255 p53 target enhancers. Interestingly, this approach also identified co-factors of p53: NFY, FOX and FOS.

Next, we applied our method to 4,106 cancer-related gene signatures from GeneSigDB[7] and MSigDB[6]. For each signature we obtained a list of motifs and regulatory tracks along with their targets that were detected for these cancer-related gene sets. From these results we created so-called "meta-targetomes". For example, we selected all signatures for which the p53 motif and the p53 ChIP-Seq track are found enriched. From all predicted p53 target regions, across signatures, we derived a meta-targetome. We validated the p53 meta-targetome using an in-house set of 801 genes up-regulated after p53 activation (in the MCF7), and found that it accurately describes p53 target genes. We used this approach for 161 different TFs and we are now using these meta-targetomes to annotate candidate cis-regulatory mutations in cancer genomes.

We are currently further validating our method and we are investigating how to use this method to identify driver mutations in the non-coding part of a cancer genome.

REFERENCES

1. Dunham, I. *et al. Nature* 489, 57–74 (2012).
2. Ernst, J. *et al. Genome research* **20,** 526–36 (2010).
3. Montgomery, S. B. *et al. Bioinformatics* **22,** 637–40 (2006).
4. Pennacchio, L. a *et al. Nature* **444,** 499–502 (2006).
5. Ferretti, V. *et al. Nucleic Acid Res* **35,** D122–6 (2007).
6. Liberzon, A. *et al. Bioinformatics* **27,** 1739–40 (2011).
7. Culhane, A. C. *et al. Nucleic Acid Res* **40,** D1060–6 (2012).

PREDICTION OF TRANSCRIPTIONAL TARGETS USING ADVANCED ENHANCER MODELS

Dmitry Svetlichnyy[1,*], *Hana Imrichova*[1] & *Stein Aerts*[1].
Laboratory of Computational Biology, University of Leuven[1], *dmitry.svetlichnyy@med.kuleuven.be*

Transcription factors are proteins that interact with DNA to regulate gene expression. However it is not well understood what distinguishes bound versus unbound genomic loci, and active versus inactive loci. Here we present an approach to select suitable training enhancers, performing selection of relevant cis-regulatory features, train enhancer models for 16 TFs, and use them for whole-genome prediction of regions bound and activated by the transcription factor.

INTRODUCTION

The binding of transcription factors (TF) to the DNA within *cis*-regulatory modules (CRM) regulates specific patterns of gene expression. TFs typically recognize small 5–15 bp DNA sequences (motifs) but only a small fraction of genomic locations matching such motifs are actually bound by the TF[1]. Furthermore, neither presence nor absence of high affinity motifs can accurately define regions of TF occupancy in the genome. Experimentally, the genome-wide occupancy of transcription factors is determined by chromatin immunoprecipitation followed by sequencing (ChIP-seq). However, observing where TFs bind in the genome does not explain the mechanisms of binding nor does it unambiguously indicate which genes are affected. Taken together, this represents a dual challenge in the study of genome control, namely (1) to predict from the DNA sequence where a TF may bind, and (2) which binding events may have functional consequences for target gene expression.

METHODS

We used a Random Forest algorithm[4] to train a classifier for 16 cancer related TFs (e.g., p53, E2F1, E2F4, STAT3, STAT5, FOXM1). We selected informative motif features, both for the query TF, and for candidate co-factors, using an in-house motif enrichment method, namely i-cisTarget2.0[3]. Once the top-scoring motif features are selected, we use a Hidden Markov Model to score all candidate sequences[5]. We analyzed the performance of the classifiers through a stratified 5-fold cross-validation procedure, and calculated both the area under the receiver operating characteristic (AuROC) and the area under the precision-recall (AuPR) curve. We compared positive sequences with negative samples, being the 200 bp-long sequence located at a distance 1 kb up and downstream from the positive CRM. Note that the training sets are imbalanced with a bias towards negative samples, as to reflect the real situation that only a small fraction of the genome is occupied by the TF.

RESULTS & DISCUSSION

To identify bona fide binding sites and target genes of a TF we applied supervised machine learning methods, taking into account both homotypic and heterotypic clusters of TF binding sites. To construct high-quality training sets with a minimum of noise we combined publicly available gene expression data for fifteen TF perturbations, with ENCODE ChIP-seq data for the same TF. We also included a sixteenth TF, namely p53, for which we have generated in-house RNA-seq and ChIP-seq data. As positive sequences we selected for each TF a high-confidence subset of *functional* target CRMs that have (i) a ChIP peak and (ii) a significant expression change of a nearby target gene.

We compared the RF model to classical models whereby the PWM of the query TF is used to score sequences with HMMs. The RF classifier yielded a much higher average AuROC of 0.92, across the 16 TFs, compared to the baseline HMM performance using PWMs only (average AuROC = 0.88). Also the AuPR of the RF classifier was higher than the baseline (0.634 versus 0.297). The best-scoring TFs were cMYC, FOXM1, ESR1 while the worst performing TFs were TAL1, NANOG, and STAT3.

Next, we used several well-performing enhancer models (with AuPR more then 0.6), namely those for p53, ESR1, cMyc, FOXM1, ATF2, E2F1, E2F4, HNF4A, ZEB1, NFE2L2 to predict TF target enhancers across the entire genome. For p53 we predicted 7393 putative binding regions in the human genome, considering a prediction with a fraction of voted trees greater then 50% as positive. We further evaluated these predictions using gene expression data and ChIP-seq data for histone modifications such as H3K27Ac. Finally, we compare the performance of enhancer models with ChIP-seq performance to identify bona fide TF target enhancers and target genes.

REFERENCES

1. Spitz F. & Furlong EEM. *Nat. Rev. Genet.* **13,** 613–626 (2012).
2. Liberzon, A. *et al. Bioinformatics.* **27,** 1739–1740 (2011).
3. Herrmann C. et al. *Nucleic Acids Res.* doi:10.1093/nar/gks543 (2012)
4. Breiman, L. Random Forests. *Mach. Learn.* **45,** 5–32 (2001).
5. Frith MC. et al. *Nucleic Acids Res.* **31,** 3666–3668 (2003).

STABLE FEATURE SELECTION TECHNIQUES FOR MICROARRAY DATA

Sofie Van Gassen[1,2,3*], Joeri Ruyssinck[1], Yvan Saeys[2,3], Tom Dhaene[1].
Department of Information Technology (INTEC)-iMinds, Ghent University, Belgium[1]; Department of Respiratory Medicine, Ghent University, Belgium[2]; VIB Inflammation Research Center, Belgium[3]. *sofie.vangassen@intec.ugent.be

Microarrays measure gene expression and these datasets are often processed with machine learning techniques to build diagnostic models. One of the major challenges is that the datasets are high-dimensional and contain a relatively low number of samples. By selecting only the most informative features in the data related to the problem at hand, the dimensionality of the dataset can be reduced, leading to an increase in performance and a reduction in complexity. Furthermore, knowing which features are important can provide vital insights in the underlying structure of the dataset and the biological processes involved. In this context, it is crucial that the selection of features is stable, meaning that small changes in the data do not lead to large changes in the selected feature set. We will do a comparative study of the stability of different feature selection techniques.

INTRODUCTION

On a microarray chip, gene expression can be measured for thousands of genes in parallel, resulting in high dimensional datasets. This may lead to complex diagnostic models, which are prone to overfitting due to the small number of available samples. To alleviate this problem, one can use feature selection techniques to reduce the dimensionality as a preprocessing step. Another advantage of applying feature selection techniques in this context is that the results can indicate which genes are most informative about the diagnostic problem. However, when these results are interpreted, attention should be paid to the stability. If small changes in the dataset result in a very different selection of the most important features, little insight can be gained.

METHODS

In microarray data, there are two main **causes for instability**[1]. Firstly, due to the small number of samples in comparison to the very large amount of features, adding or removing one sample might have a large impact on the results. Secondly, many features are often strongly related in microarray data. As such, multiple feature subsets can give the same kind of information and equally good results. If feature selection techniques will randomly select one of the possible subsets, this strongly influences their stability.

We will compare the results of **four commonly used feature selection techniques**: Pearson correlation, Support Vector Machines, Relief and the Elastic Net. All of them output feature weights as a result, which we then convert to feature subsets. To measure the stability of a method, we use the **Kuncheva measure**[2].

To increase the stability of the basic feature selection techniques, we use **ensembles**[3]. We create an ensemble by training a collection of feature selectors on different bootstraps of the data. In the end, the results of the feature selectors are aggregated to a single result. By combining different results to an average, we are able to get more stable results.

Feature clustering is a technique that can be used as a basis for feature selection, but also helps to gain more insight in the structure of the data. K-means clustering is a very simple clustering technique. It is limited because it will only be able to make spherical clusters. **Dense Feature Groups**[4] is a clustering technique based on the mean shift procedure. This technique clusters features based on how densely they are placed together and is not limited to spherical clusters. We also introduce a novel feature selection technique based on **hierarchical clustering**. There might be some variations in the initial clustering of the features, but the results will get more stable once more features are gathered together. We test this method for different distance measures between clusters.

RESULTS & DISCUSSION

Tests show that for the basic techniques, a simple technique as computing the correlation might give more stable results than the more advanced techniques such as Relief or Elastic Net, because it does not take into account which features are related.

For less stable techniques, ensembles provide a good way to improve their stability (Figure 1). However, this happens at the cost of a longer processing time.

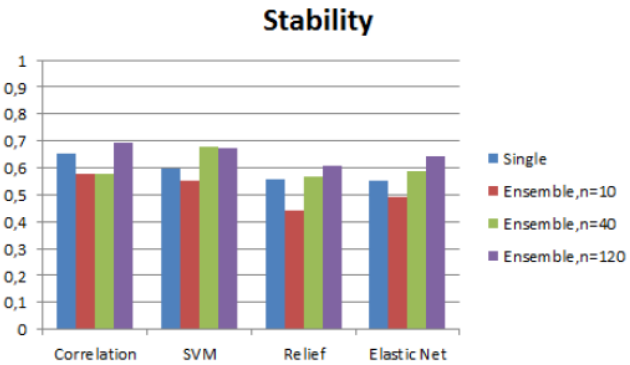


***FIGURE 1**. Stability results for ensembles with a varying number of bootstraps.*

Finally, in our search for a stable cluster algorithm, dense feature groups and hierarchical clusters with single linkage give stable and accurate results.

REFERENCES

1. He Z & Yu W. *Comput Biol Chem* **34**, 215–225 (2010).
2. Kuncheva LI. *Artificial Intelligence and Applications*, 421–427 (2007).
3. Saeys Y. et al. *Machine Learning and Knowledge Discovery in Databases*, 313-325 (2008).
4. Yu L. et al. *Proceedings of the 14th ACM SIGKDD international conference on Knowledge discovery and data mining,* 803-811 (2008).

EXPERIMENT-SPECIFIC PROBE SET ANNOTATION FOR AFFYMETRIX GENE EXPRESSION DATA

Tatsiana Khamiakova[1],*, *Ziv Shkedy*[1], *Hinrich Göhlmann*[2], *Willem Talloen*[2] , *Adetayo Kasim*[3].
Interuniversity Institute for Biostatistics and Statistical Bioinformatics, CenStat, Universiteit Hassel, Belgium[1]*; Janssen, Pharmaceutical Companies of Johnson and Johnson, Belgium*[2]*; Wolfson Research Institute for Health and Wellbeing, Durham University,UK*[3]*. Tatsiana.khamiakova@uhasselt.be*

The probe-level analysis, filtering and summarization methods of Affymetrix microarray data largely rely on the correctness of a probe set corresponding to a gene. This assumption can be violated due to a biological phenomenon known as differential splicing or due to the technical problems. In this work, we illustrate how the probe-level linear models can be used in order to improve the filtering and how to detect groups of the probes within a probe set, which we term the informative core of the probe set that capture the transcript or its part of a given gene. Furthermore, we illustrate how patterns in the data which are not expected to appear in the Platinum spiked data set can be detected by the extended mixed model.

INTRODUCTION

The definition of a probe set for which groups of probes are mapped to the same transcript target has been a source of concern in Affymetrix GeneChip data, since it is essential for summarization, analysis, and interpretation of the results in a microarray experiment[1]. When more than a half of probes in a probe set fail to pick up the transcript either due to the alternative splicing or due to some technical problems, the whole probe set may be discarded from further analysis[2]. However, it may still carry the important information for the part of a transcript, which has been captured.

The aim of this work is to illustrate (1) how to classify probes in a probe set, and (2) how to use linear mixed-effects model to discover probe sets with technical or biological problems on the probe-level.

METHODS

Throughout this section Y_{ij}denotes a log(PM) measurement for a probe j on the array i. All models discussed below are fitted per probe set.

Classification of probes in probe sets

A data-driven way of identification of subgroups in a probe set by using a mixture model was proposed in [1]. The model includes G classes of probes in a probe set and is formulated as

$$Y_{ij} = \sum_{s=1}^{G} \pi_s N(\mu_j + b_{is}, \sigma_\varepsilon),$$

where π_s, $s = 1..G$, is the mixing probability for a component s and b_{is} are array-specific random effects for the component s. The parameters of the model are estimated within Bayesian framework by using Gibbs sampler with the normal and Gamma priors for the unknown parameters.

Probe-level linear mixed model

The basic probe-level linear model used for filtering[2], ignoring grouping of probes is given by

$$Y_{ij} = \mu_j + b_i + \varepsilon_{ijs}, i = 1..n \ , j = 1..l, \qquad (1)$$

where $b_i \sim N(0, \sigma_b^2)$ is the random effect of an array, and $\varepsilon_{ij} \sim N(0, \sigma_\varepsilon^2)$ is a residual error.

The probe-level linear mixed model, which takes into account probes labels, is formulated as

$$Y_{ijs} = \mu_j + b_{is} + \varepsilon_{ijs}, i = 1..n_s, j = 1..l, s = 1..G,$$

where $b_{is} \sim N(0, \sigma_{bs}^2)$ is the stratified random effect of an array, σ_{bs}^2 is the variance of a component s , and $\varepsilon_{ijs} \sim N(0, \sigma_{\varepsilon s}^2)$ is a residual error with a component-specific variance to account for heteroscedasticity within each component.

The within-component intra-class correlation (ICC) ρ_s is given by $\rho_s = \sigma_{bs}^2/(\sigma_{bs}^2 + \sigma_{\varepsilon s}^2)$.

At the next step, for each probe set, the ICC from basic model (2) $\rho_0 = \sigma_b^2/(\sigma_b^2 + \sigma_\varepsilon^2)$ is compared to each ρ_s and in case when $\rho_s > \rho_0$, we can re-define a probe set as follows: in the setting with two components, where one component has $\rho_{s1} > 0.5$, and the other with low $\rho_{s2} < 0.5$, we can consider excluding component s2 from downstream analysis or label as a non-informative part of a probe set.

Data

The method has been applied to the Platinum Spike data[3] with 18,952 probe sets with 22 probes on a Drosophila GeneChip. The probe-level data have been background corrected and normalized by RMA[4].

RESULTS & DISCUSSION

Based on the posterior probabilities of mixture model, 17,051 out of 18,952 probe sets had one component, 1,860 probe sets had two components and 41 probe sets had three components. The detection of probe sets with two or three components due to expression values can be explained by technical artefacts, like unspecific binding or inefficiencies in probe synthesis for some of the probes. There were more informative probe sets detected, by taking into account informative components, than the informative probe sets in the basic model (3,640 vs. 2,461, respectively). By checking probe sets, which became informative after grouping was taken into account, we have observed that only a subset of a probe set measured the spiked transcript or cross-hybridization took place. The proposed method can be applied to any gene expression data set, profiled by GeneChip, to identify both technical artefacts and the alternative isoforms of genes. Furthermore, it can be extended to exon and GeneST arrays.

REFERENCES

1. Kasim et al. *unpublished* (2012)
2. Kasim et al., *Stat Appl Gen and Mol Biol* **9** (1)(2010)
3. Zhu et al. *BMC Bioinformatics* **11**, 285 (2010).
4. Irizarry et al. *Biostat*, **4**, 249-264 (2003).

BI-CLUSTERING GENE EXPRESSION DATA UNDER CONSTRAINTS

Thanh Le Van[1,*], *Ana Carolina Fierro*[2], *Tias Guns*[1], *Matthijs van Leeuwen*[1], *Siegfried Nijssen*[1], *Luc De Raedt*[1], *and Kathleen Marchal*[2,3].

Depts. of Computer Sciences[1], *and Microbial and Molecular Systems*[2], *KU Leuven ; Dept. of Plant Biotechnology and Bioinformatics*[3], *Ghent University.* [*]*thanh.levan@cs.kuleuven.be*

This paper presents a constraint-based approach to mining bi-clusters in gene expression data. Instead of designing an algorithm for each specific task, we propose to use constraint programming to turn the mining problem into a constraint satisfaction and/or optimisation problem. We demonstrate this promising approach on two cases. The first is to mine a single constant-row bi-cluster under noise constraints. The second is to mine a set of generic noisy constant-row bi-clusters under structure constraints, which is called a staircase pattern.

INTRODUCTION

In gene expression analysis, we are given a data matrix in which rows correspond to genes, columns correspond to conditions and data shows expression values of genes in conditions. A bi-clustering algorithm typically finds a subset of genes that shows an approximately constant value for a subset of conditions. The submatrix formed by the selected subset of genes and conditions is called a bi-cluster. Bi-clusters are interesting as the relationship between conditions and genes provides insight in the correlation of genes and can be used for finding perturbed biological processes or predicting gene regulation networks. The challenge that we study is to develop a generic and extendible approach to take into account requirements of a good bi-cluster. Some desirable properties include: rows need to be approximately constant; bi-clusters do not have much noise; constraints can be easily added or removed when we want to perform integrative data analysis.

METHODS

Different from earlier approaches, we propose a more general way to formalize and solve the problem using constraints. The prominent contribution lies in the fact that the entire model, which consists of an objective function and a set of constraints, is specified in a declarative programming language and solved using existing techniques supported by the language. In practice, we chose the constraint programming (CP) paradigm for modelling and solving. Working in this way, we have a number of advantages. First, it is a declarative approach. We can exploit built-in solving capabilities implemented by CP solvers, for instance constraint propagation and search strategies, to avoid re-inventing the wheel for common tasks and have more time to focus on modelling. The program we build will be easier to maintain or extend. Second, it is not hard to extend the model to other settings, for example, detected bi-clusters should have consistent patterns in another data matrix.

We select two settings for the mining task to present the proposed methodology.

In the first setting, we demonstrate how to compile the problem of mining a single fault-tolerant constant-row bi-cluster that covers the largest part of the data into a constraint optimisation problem. We also show how to use large neighbourhood search, a type of local search, to approximate the optimal solution.

In the second setting, we illustrate the extensibility of the framework to pattern set mining under structure constraints. More precisely, we want to find a set of fault-tolerant constant-row bi-clusters which resembles a staircase[1]. As the quality of the staircase depends on the user-defined noise thresholds, we propose a two-phased mining approach. First, we generate staircase candidates by solving a number of constraint satisfaction problems. Then, we use the Minimum Description Length (MDL) principle to select the best one. According to the MDL principle, the best model is the one that compresses the data best. In this case, a model is a staircase.

RESULTS & DISCUSSION

We experimented with a number of synthetic data sets of 1000 rows and 120 columns with varying noise levels and a number of staircase steps. Figure 1 shows that our model can recover most of the staircase and the MDL scores help us to select the best candidate.

In real data sets, we encountered the scalability problem of solvers. Besides that, detected staircases do not often have discernible steps (high noise outside). In the future, we plan to integrate with more data sets to increase the quality of the bi-clusters.

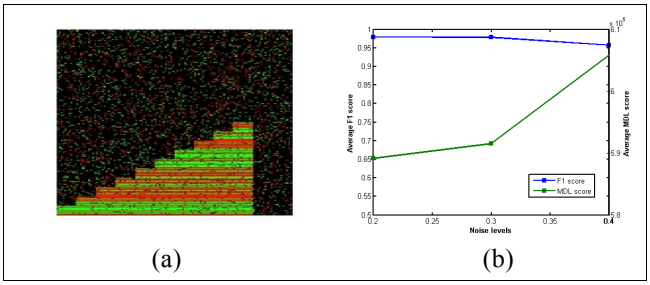


***FIGURE 1**. a) A synthetic staircase. b) Relating F1-scores to MDL-scores*

REFERENCES

1 Le Van, T., Fierro Gutiérrez, A., Guns, T., van Leeuwen, M., Nijssen, S., De Raedt, L., Marchal, K. (2012). Mining local staircase patterns in noisy data. *12th IEEE International Conference on Data Mining Workshops*. International workshop on Co-Clustering and Applications (CoClus'12) in conjunction with IEEE ICDM 2012.

GALAHAD – A WEB SERVER FOR GENE EXPRESSION DATA ANALYSIS IN SUPPORT OF DRUG DEVELOPMENT

Griet Laenen[1,2,*], *Amin Ardeshirdavani*[1,2], *Yves Moreau*[1,2] & *Lieven Thorrez*[1,3].

Dept. of Electrical Engineering (ESAT), STADIUS Center for Dynamical Systems, Signal Processing and Data Analytics, KU Leuven[1] ; *iMinds Future Health Dept., KU Leuven*[2] ; *Dept. of Development and Regeneration @ Kulak, KU Leuven*[3].

[*]*griet.laenen@esat.kuleuven.be*

With many candidate drugs failing in the late clinical stages, drug development is a high-risk business. In order to reduce the high rates of attrition, a better knowledge of a drug's mechanism of action is required. Hence we have developed Galahad, a web server for the analysis of gene expression data following drug treatment, aimed at predicting a drug's molecular targets and physiological effects.

INTRODUCTION

The pharmaceutical industry is facing unprecedented productivity challenges. Attrition rates have risen sharply, especially in late-phase clinical trials. With safety and efficacy being the main bottlenecks, a better knowledge of a candidate drug's mode of action and its off-target effects could be of substantial value to drug development. DNA microarray technology enables a genome-wide analysis on the transcriptional response to a compound treatment, and thus can provide valuable information for identifying the compound-protein interactions and resulting effects prior to clinical trials. In addition, this information may also be useful for already marketed drugs, in the light of drug repositioning.

METHODS

We have developed a new, easy-to-use web server called Galahad, for the in-depth exploration of a drug's mode of effect based on gene expression changes following treatment. Our software provides the main tools needed for gaining new insights into the biological effects of a drug by combining

- **preprocessing** of gene expression data obtained from different Affymetrix array types;
- **quality assessment and exploratory analysis** of these data to ascertain data quality, uncover experimental issues or sample mix-us, and help in deciding whether certain arrays need to be considered as outlying;
- **differential expression analysis** to determine the significance of gene up- or down-regulation following drug treatment by fitting a linear model to the expression data for each gene;
- genome-wide **drug target prioritization** by means of an in-house developed algorithm for network neighborhood analysis integrating the expression data with functional protein association information[1];
- prediction of **Reactome pathways** involved in the drug's mode of effect;
- identification of associated **disease phenotypes** from the Human Phenotype Ontology enabling side effect prediction and drug repositioning.

RESULTS & DISCUSSION

All of the above functionalities can be demonstrated on gene expression data for treatment with drugs exhibiting a well-defined mechanism of action. One such drug is infliximab, a tumor necrosis factor (TNF)-binding monoclonal antibody marketed under the brand name Remicade and used in the treatment of several autoimmune diseases. Infliximab target prioritization based on the gene expression profiles from eleven Crohn's colitis patients treated with this drug[2] ranks the target TNF in the top 1%, although not differentially expressed following treatment. Enrichment analysis on the significantly up- and down-regulated genes returned several immune pathways, as well as links to disease phenotypes reported in literature either as a known indication, a side effect, or a possibility for repositioning. By application to a larger set of well-characterized chemical drugs, Galahad will now be further optimized.

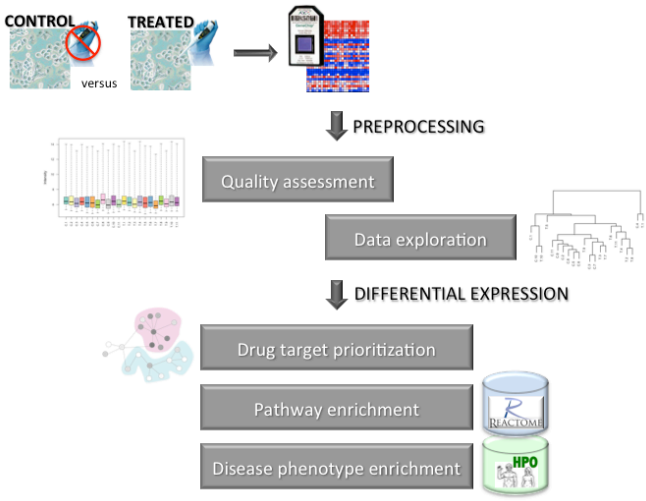


***FIGURE 1**. Overview of analyses provided by Galahad.*

REFERENCES

1. Laenen G. *et al. Mol BioSyst* **9**, 1676-1685 (2013).
2. Arijs I. *et al. Am J Gastroenterol* **106**, 748-761 (2011).

COMPARISON OF METHODS FOR PATTERN RECOGNITION IN TOXICOGENOMICS TIME SERIES

Diana M Hendrickx [1,*], *Danyel G J Jennen*[1], *Jacob J Briede*[1], *Rachel Cavill*[1], *Theo M de Kok*[1] & *Jos C S Kleinjans*[1].

Dept. of Toxicogenomics, Maastricht University, The Netherlands[1]. [*] *d.hendrickx@maastrichtuniversity.nl*

During the last decade, the field of toxicogenomics has grown enormously due to the need for alternatives to animal testing for chemical safety assessment. To discover mechanisms of action of a toxic compound, it's important to establish relationships between gene expression and phenotypic endpoints (e.g. formation of DNA adducts, DNA damage, protein oxidation, apoptosis). This is one of the goals of pattern recognition methods. In this study, pattern recognition methods for the identification of genes whose expression time-courses are associated with phenotypic endpoints are compared. The results show that none of the methods can identify all gene-phenotypic endpoint relationships and methods need to be combined.

INTRODUCTION

Traditional toxicity testing consists of animal experiments. Due to concerns about animal welfare and reliability of animal experiments for studying human toxicity, research on the potential application of toxicogenomics approaches as an alternative to animal testing has expanded considerably. Gene-phenotype relationships are important for the discovery of mechanisms of action of a toxic compound. In this study, methods for associating time-courses of gene expression with time-courses of phenotypic endpoints are compared.

METHODS

Pattern recognition methods for time-courses have been shown to be very effective in answering many biological questions, because biological entities (e.g. genes) belonging to the same cluster are assumed to be functionally related[1,2]. However, studies comparing different algorithms for pattern recognition are lacking[1].

To perform a comparative study of pattern recognition methods for identification of gene-phenotypic endpoint relationships from time-courses, four representative methods were chosen: k-means clustering, short time series expression miner (STEM), linear mixed model (LMM) mixtures and dynamic time warping (DTW4omics). STEM clusters time-courses by assigning them to predefined profiles which are significantly present in the data[3]. An LMM mixture is a model-based clustering approach that takes into account variability between replicates[4]. DTW4omics is a pattern recognition method for time-courses that takes into account that similar patterns do not always occur simultaneously[5]. K-means divides a dataset into clusters so that each observation is assigned to the cluster with the nearest center[1]. The four methods were applied on two published data sets on HepG2 cells: the response to benzo(a)pyrene by van Delft *et al* (2010)[6] and menadione by Deferme *et al* (2013)[7]. Both data sets consist of time-series of both gene expression (Agilent and Affymetrix microarrays respectively) and phenotypic endpoints.

Lists of genes whose time-courses are associated with time-courses of an endpoint are identified and the overlap among the four methods determined. Both positive and negative association are considered. Gene Ontology (GO) and pathway analysis are performed with ConsensusPathDB[8]. Both GO and pathway lists are compared among the four methods and with GO and pathway lists determined with text mining tools. The clusters found by the different methods are visualized with Cytoscape.

RESULTS & DISCUSSION

Gene, pathway and GO lists show low overlap among the methods (for an example for GO lists, see Figure 1), which suggests that the methods need to be combined to get a more complete view on the relationships between genes and phenotypic endpoints. GO and pathway lists generated by the methods show a large overlap with GO and pathway lists found with text mining tools. Additionally, relationships not occurring in the literature represent new hypotheses. Testing these hypotheses experimentally can lead to the discovery of new mechanisms of action for a toxic compound.

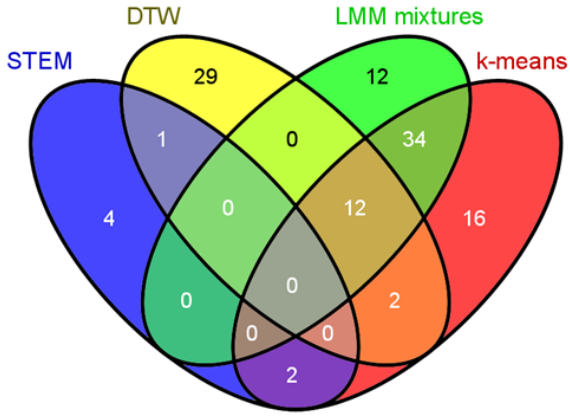


***FIGURE 1**. Response of HepG2 to benzo(a)pyrene. GO analysis applied on gene lists positively associated with DNA adducts. Number of GO for each pattern recognition method and overlap among the methods.*

REFERENCES

1. Liao TW *Pattern Recognition* **38**, 1857-1874 (2005).
2. Bar-Joseph Z *Bioinformatics* **20**(16), 2493-2503 (2004).
3. Ernst J & Bar-Joseph Z *BMC Bioinformatics* **7**:191 (2006).
4. Celeux G *et al. Statistical Modelling* **5**, 243-267 (2005).
5. Cavill R *et al. PLoS One* **8**(8):e71823 (2013).
6. van Delft JH *et al. Toxicol Sci* **117**, 381-392 (2010).
7. Deferme L *et al. Toxicology* **306**, 24-34 (2013).
8. Kamburov A *et al. Nucleic Acids Res* **41**, D793-D800 (2013).

UNVEILING THE MECHANISMS OF ACTION AND THE SIDE EFFECTS OF DRUGS BY COMPARATIVE MODULE ANALYSIS

Daniele Pepe[1,2,*] *and Yves Moreau* [1,3].

Dept of Electrical Engineering, Esat-SCD, K.U. Leuven,3001 Leuven, Belgium,[1] *Dept of Brain and Behavioral Sciences, Medical and Genomic Statistics Unit, University of Pavia, Italy;* [2], *Dept. IBBT-K.U.Leuven Future Health Department, 3001 Leuven, Belgium*[3]. *Daniele.Pepe@esat.kuleuven.be*

The understanding of the molecular mechanisms beyond the side effects and the mode of action of drugs, is of utmost importance in medicine and in pharmacology. Many approaches were proposed as the analysis of chemical structures, text mining and the individuation of gene expression profiles. However, a new paradigm has been emerged recently that has revolutionized the world of the biology and the medicine: complex phenomena in nature are the result of the connection among different components. Following the principles of the network medicine, we propose a method that allows unveiling drug side effects and mode of action by the comparison of drug and disease genic modules.

INTRODUCTION

As well illustrated by Barabasi et al.[1], a disease is the consequence of the perturbation of molecular component connections. The role of a drug should be that to restore the healthy state of the patient. Many times the effect of a drug brings side effects. The comparison of the molecular network on which a drug acts with the disease molecular module could reveal important information to predict side effects and the mode of actions of the drug. Using gene expression data analysis, based on pathway analysis and structural equation modeling (SEM)[2], it has been possible to find drug and disease modules and then to compare them.

METHODS

For each microarray dataset the following steps were performed:

1. differential gene expression analysis, finalized to find differential expressed genes (DEGs)[3] and perturbed pathways[4];
2. individuation of the shortest paths that connect every couple of DEGs;
3. individuation of the perturbed paths using the multiple-group SEM analysis;
4. merging of all significant shortest paths to have the drug or the disease module.

The method is similar to that illustrated from Pepe et al.[5]. To evaluate the goodness of the gene selection, a disease enrichment analysis[6] on the list of the module genes was used. Then a comparison of the modules, by the fold change of DEGs and the enrichment analysis, was performed. The procedure was tested on two microarray datasets downloaded from GEO[7]: one relative to the effect of infliximab on patients affected by ulcerative colitis (UC) (id GSE 23597) and the other considers the gene expression in UC patients against healthy samples. (id GSE 23597)

RESULTS & DISCUSSION

The differential analysis has revealed 306 DEGs for infliximab data and 794 DEGs for UC data. On each set of genes a pathway analysis was performed using the algorithm SPIA[4]. The results were very interesting considering that, the two sets of DEGs shared many important KEGG pathways as the chemokine signaling and cytokine-cytokine receptor interaction pathway. The perturbed pathways for each experiment were merged to create a unique networks (1112 nodes and 3920 edges for infliximab and 908 nodes and 3613 edges for UC). Then for each network a list of shortest paths between DEGs was generated and tested with multiple group SEM. The significant shortest paths were merged to obtain the drug (91 nodes and 180 edges) and disease modules. (51 nodels and 92 edges). The disease enrichment analysis for the nodes of each module has showed the goodness of the procedure considering that, diseases as UC, rheumatoid arthritis (RA), lupus erythematosus (SLE) has been resulted enriched. The comparison of DEGs, in the two modules allowed understanding the mechanisms of actions and the side effects of the drug. In fact, the DEGs present only in the drug module, has enriched diseases as SLE, a side effect of infliximab and the common DEGs showed opposite fold changes.

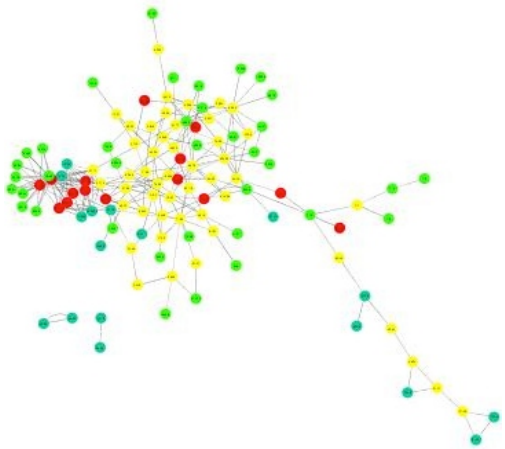


FIGURE 1*. Green nodes are DEGs in common, light blue nodes are DEGs only in IF data, red nodes are DEGs only in UC data, yellow nodes are not DEGs that connect the DEGs.*

The procedure, starting from the paradigm of network medicine, permitted to define and to compare drug and disease modules, finalized to understand the effects of drug from a molecular network point of view.

REFERENCES

1. Barabási A. et al. *Nat Rev Genet*,**12**, 56-68 (2011).
2. Bollen K. A. *John Wiley & Sons, Ltd*. (1998).
3. Tusher V. G et al. *PNAS*, **98**, 5116-5121 (2001).
4. Tarca, A. L. et al. *Bioinformatics*, **25**, 75-82. (2009).
5. Pepe, et al. *Advances in Latent Variables* (2013).
6. Yu, G., & Wang, L. G. (2012).
7. Edgar, R et al. *Nucleic Acid Res*, **30**, 207-210 (2002).

PROTEIN IDENTIFICATION BASED ON RIBOSOME TARGETED MRNA FRAGMENTS

Jeroen Crappé[1], Alexander Koch[1], Elvis Ndah[1], Sandra Steyaert[2], Petra V. Damme[2], Gerben Menschaert[1,].*

*Lab of Bioinformatics and Computational Genomics (BioBix), Department of Mathematical modelling, Statistics and Bioinformatics, Faculty of Bioscience Engineering, Universiteit Gent, B-9000, Gent. *gerben.menschaert@ugent.be*

The newly developed ribosomes profiling (RIBO-seq) approach provides genome-wide information about protein synthesis by monitoring mRNA that enters the translation machinery, while highly sensitive mass spectrometry provides information about the protein composition of a sample. Integrating these technologies could provide more intuitive information about the protein molecules being synthesized and the identification of novel translation products as well as a better understanding of the translation mechanism.

INTRODUCTION

The recently developed ribosome profiling approach based on deep sequencing of ribosome protected mRNA fragments that monitors translation at the codon level has revolutionized the study of translation mechanism[1]. It allows mapping of the location of translating ribosome on mRNA with sub-codon precision[2]. RIBO-seq profiling may indicate which portion of the genome is actually being translated at the time of the experiment as well as account for sequence variations such as single nucleotide polymorphism, indels and RNA splicing. Integrating a custom database based on RIBO-seq predicted sequences, including novel and unexplored protein isoforms into Swiss-Prot may improve the search space for MS-based proteomic studies as Swiss-Prot alone does not include these non-annotated alternative translation products[3].

We present a proteogenomic pipeline that takes advantage of all extra information extracted from RIBO-seq/RNA-seq data, to derive an optimized protein search database. The database will constitute a combination of RIBO-seq derived translation products and UniProtKB/Swiss-Prot.

METHODS

To investigate the impact of combining RIBO-seq predicted translation sequences with Swiss-Prot we use matching RIBO-seq[4], gel-free shotgun and N-terminal COFRADIC proteomic data from mouse embryonic stem cells (mESC).

Figure 1 depicts the work-flow diagram of the bioinformatics pipeline used to construct a customized database of translation products based on the Ensembl annotation.

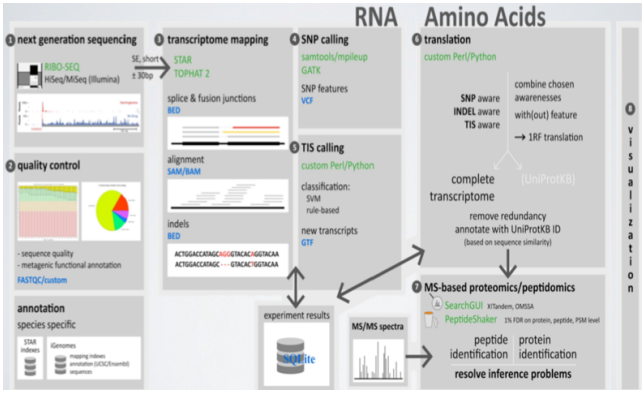


***FIGURE 1**. Proteogenomic pipeline.*

Our pipeline consists of 8 major steps including transcription mapping by STAR or TopHat2, SNP and indel calling using SamTools-mpileup or GATK, translation initiation site (TIS) calling using rule-based or SVM training, and translation product assembly taking into account all the sequence variation information obtained from the RIBO-seq data. The final step constitutes combining the transcript database with Swiss-Prot for MS-based proteomics/peptidomics.

RESULTS & DISCUSSION

Incorporating RIBO-seq derived translation products obtained from deep sequencing of ribosome protected fragments provides a means to capture translation products that would have otherwise been missed by Swiss-prot search alone. Using the combined data set we were able to identify 259 non-annotated translation start sites, indicating alternative or wrongly annotated protein translation sites as well as 16 N-terminal extended protein forms and four translated uORFs. Together with the identification of new protein splice variants and proteins including SNPs, characterization of these new translation products revealed the use of multiple near-cognate (non-AUG) start codons[5]. This demonstrates an increase in overall protein identification rate as compared to only searching UniProtKB/Swiss-Prot.

The integration of deep sequencing and mass spectrometry will be instrumental in the study of translation mechanism and the identification of novel transcripts. It might reveal in-depth information about the yet incompletely understood quantitative nature of translation initiation and its associated co-translational N-terminal protein modifications (e.g., protein N-terminal acetylation) as well as the identification of small translation products[5].

REFERENCES

1. Michel, M *et al. WIREs RNA* (2013).
2. Lee S *et al. Preceeding of Nat. Acad. Of Sci USA* 109, E2424-32 (2012).
3. Helsens K *et al Journal of Proteomics research* 3578-3589 (2011).
4. Ingolia, *N. et al. Cell* 147, 789-802 (2011).
5. Gerben, M *et al Mol Cell Proteomics*, 1-41 (2013).

PROBIC-II: SIMULTANEOUSLY DETECTING COEXPRESSION MODULES AND THEIR REGULATORY PATTERNS

Yan Wu [1], Lieven Verbeke [2], Carolina Fierro [1], Jan Fostier [2], Kathleen Marchal [1,2,3]*
Department of Microbial and Molecular Systems , KU Leuven, Belgium[1]; Department of Information Technology, Ghent University, Belgium[2]; Department of Plant Biotechnology and Bioinformatics, Ghent University, Belgium[3].
**yan.wu@biw.kuleuven.be*

ProBic-II is a probabilistic framework to integrate publicly available microarray and regulatory motif data. It searches for regulatory modules consisting of co-regulated genes, the conditions under which the genes are co-expressed and the motifs shared by these genes.

INTRODUCTION

Approaches to explore transcriptional regulation are customarily based on clustering/biclustering to search for condition-dependent coexpression, followed by motif detection. However, a co-expression module is ill defined as its size in the number of genes is parameter-dependent. Thus, we developed ProBic-II, a tool that simultaneously searches for co-expression modules and their regulatory patterns.

METHODS

ProBic-II is based on the Probabilistic Relation Model (PRM), which has the advantage of synthesizing Bayesian network together with relational domain. ProBic-II optimizes the combined task of learning co-expressed genes and their common motifs in an iterative way through an Expectation-Maximization (EM) based strategy.

We start for each regulatory module with a set of seed genes.

Co-expression Bicluster step: ProBic-II identified a set of genes tightly coexpressed with the seed set, gene Group A (GA). This step is performed within the classes of Gene, Condition and Expression levels in the ProBic-II model.

Co-regulation Pattern step: For the highly co-expressed genes in GA, we search for the motif that is most representative for the gene set. The latter is based on the average score of the motif instances found in the intergenics of the genes in the gene set. The most representative motif will be the highest scoring one. Genes other than the ones already in the gene set for which the intergenics have a motif score higher than the average score of the motif in the already selected gene set will be added to the geneset, resulting in the extended gene Group B (GB) and its chosen motif.

Iteration step: If GB is not identical to GA, the GB genes are used as seeds again in a novel Co-expression Biclustering step. These alternate steps between coexpression biclustering and motif detection end when the gene list is stable both in the coexpression as in the motif space.

Fuzzy Clustering step: Fuzzy Clustering is used to filter calculate an ensemble module from overlapping modules.

RESULTS & DISCUSSION

To assess the performance of ProBic-II we performed a benchmark analysis. We utilize ProBic-II on a large scale *Escherichi coli* (*E.coli*) compendium together with regulatory motif data obtained by screening the whole genome with known motifs extracted from RegulonDB . We used as seed all genes from known regulons in RegulonDB. For each regulon we assessed the extent to which we could recover the full regulon (recall) and the number of additionally recruited genes (Figure 1). ProBic-II tends to more accurately recapitulate regulons of local regulators than those of global.

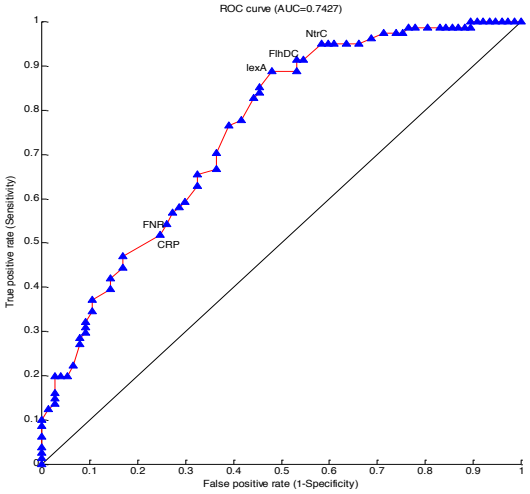


***FIGURE 1**. ROC plot: recall of regulon versus the fraction of novel predictions of regulon*

Based on the performance above, we applied ProBic-II on *Salmonella typhimurium LT2* (*S. typhimurium LT2*), which is closely related to E.coliI. By reconstructing the regulon of S. typhimurium LT2, and further comparing the regulations of both homologous genes, we will retrieve those genes, that are integrated into new pathways or new interactions increasing the probability of becoming involved in a vital evolutional process.

REFERENCES

1. Lemmens,K. *et al. Genome Biol.*, 10, R27(2009).
2. Segal,E. *et al. Nat. Genet.*, 34, 166–176(2003).
3. Hui,Z.*et al. BMC Bioinformatics*. 12(Suppl 1): S37(2011).
4. Beer,M.A. and Tavazoie,S. *Cell*, 117,185–198(2004).
5. Ruby,J.G. *et al. Genome Res.*, 17, 1850–1864(2007).
6. Bar-Joseph Z,et al.Nat Biotechnol, 21:1337-1342(2003).
7. M. Claeys.et al.*Bioinformatics,*28 (14): 1931-1932 (2012).

MECHANISTIC INTERPRETATION OF GENE LISTS USING INTERACTION NETWORKS

Dries De Maeyer[1,*], Joris Renkens[2], Luc De Raedt[2] & Kathleen Marchal[1,3,4].
Dept. of Microbial and Molecular Systems, K.U.Leuven[1], Dept. of Computer Science, K.U.Leuven[2], Dept. of Plant Biotechnology and Bioinformatics, Ghent University[3], Dept. of Information Technology, IMinds, Ghent University[4]

High-throughput experiments become a standard in current wet-lab experiments to study phenotypes yielding large lists of genes. Classicaly these lists have been analysed using enrichment methods, however these methods have the limitation that they do not provide a mechanistic insight into the pathways/processes. Therefore we propose and illustrate the use interaction networks, in combination with sub-network selection algortims such as PheNetic (De Maeyer et al., 2013), as a new analysis tool for the interpretation of these gene lists.

INTRODUCTION

Functional gene list interpretation is becoming more and more important as high-throughput experiments are becoming a standard in wet -lab practice. Currently, this interpretation is mainly performed using statistical enrichment analysis (Kathri et al., 2012). The results of these analysis methods are lists of processes and pathways that could play a role in the phenotype under study. However, there is no global picture or mechanistic representation of the processes and pathways that drive the actual phenotype. Interaction networks representing the knowledge about the interactome of an organism allow to overcome this limitation. By selecting *in silico* the active parts of the interaction network for a specific phenotype using high-throughput results, we can gain an integrated and mechanistic insight into the workings of a specific phenotype.

METHODS

Interaction networks are a representation of the interactome of an organism. These networks integrate multiple layers of interactions compiled from the vast amount of public knowledge. Based on the degree of evidence each interaction is assigned a probability representing the degree of belief in the underlying interaction. Assigning this probability allows to add vast amounts of uncertain information to the network such as information from text-mining tools and high-throughput experiments.

PheNetic (De Maeyer et al., 2013) is a sub-network selection framework that allows to identify the active part of these interaction networks in a two-step procedure. In a first step it samples the most probable paths between activated genes in the network. In a second step it selects a sub-network from the interaction network by selecting from the sampled paths specific interactions which contribute to most paths, have the highest degree of belief and best link to the activated genes. This resulting sub-network represents the mechanism predicted to drive the phenotype under study.

RESULTS & DISCUSSION

To prove the potential of PheNetic it was applied to reanalyse a KO dataset associated with acid resistance in *E. coli* (De Maeyer et al., 2013). This proof of concept of our method allowed to identify active sub-networks which were shown to represent the previously identified mechanisms and regulators behind acid resistance in *E. coli*. Additionally we show the results of an analysis of time-series expression dataset obtained from *S. typhymurium* biofilms (Van Puyvelde et al., submitted). Using sub-network selection algorithms the activated sub-networks between different time points in planktonic phase and biofilm were selected, allowing a visual and integrated analysis of the processes involved in biofilm formation.

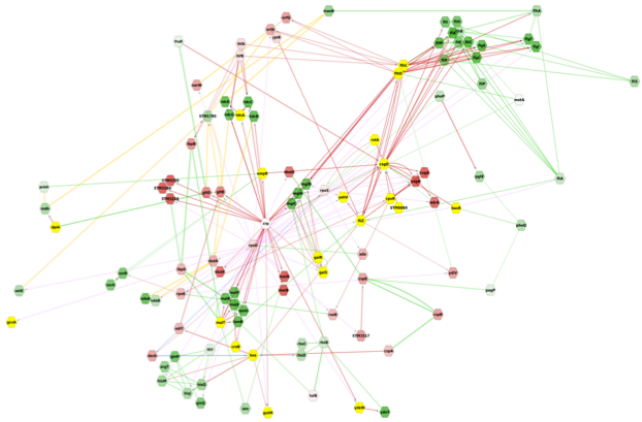


***FIGURE 1**. Active sub-network selected by PheNetic associated with early biofilm development in S. Typhymurium.*

Based on the practical applications we can conclude that interaction networks in combination with sub-network selection algorithms have a large potential in the interpretation of gene lists resulting from "omics" experiments. The framework developed for PheNetic is a flexible and configurable tool that can be used on diverse types of networks for different types of data.

REFERENCES

1. De Maeyer, et al. PheNetic: Network-based interpretation of unstructured gene lists in E. coli. Molecular BioSystems. doi:10.1039/c3mb25551d (2013).
2. Khatri, et al. Ten years of pathway analysis: current approaches and outstanding challenges. *PLoS computational biology*, *8*(2), e1002375. doi:10.1371/journal.pcbi.1002375 (2012).

INFERRING THE DIRECTION OF GENE INTERACTIONS

Miguel Lopes[1,*], Gianluca Bontempi[1]
Machine Learning Group, Dept. of Computer Sciences[1], Université Libre de Bruxelles
Interuniversity Institute of Bioinformatics of Brussels

Current approaches to infer gene regulatory networks from time series are based on the estimation of the dependence or conditional dependence between lagged variables. We assess the importance of taking into account dependencies occurring at multiple lags and also of auto-correlation, using the Granger causality framework.

INTRODUCTION

The inference of gene regulatory networks from time series is usually based on lagged regression coefficients, correlations or partial correlations. Current state of the art approaches estimate the dependence between variables lagged by only one time point, which is often 1. The consideration of auto-correlation is also not generalized. We inferred the direction of gene interactions, from gene expression time series, based on the notion of Granger causality. The impact of considering auto-correlation and dependencies at multiple time points is investigated.

METHODS

Let Y_t and X_t be random variables representing the expression of genes Y and X, at time t. Y_t can be modelled as a function of the past values of both genes:

$$Y_t = \alpha_0 + \left(\sum_{n=1}^{L} \alpha_n Y_{t-n}\right) + \left(\sum_{n=1}^{L} \beta_n X_{t-n}\right) + \epsilon_{1t} \quad (1)$$

If Y_t does not depend on previous values of X (Granger non-causality from X to Y), β_n=0, and equation (1) reduces to:

$$Y_t = \alpha_0 + \left(\sum_{n=1}^{L} \alpha_n Y_{t-n}\right) + \epsilon_{2t} \quad (2)$$

Let RSS1 and RSS2 be the residual sum of squares of equations (1) and (2), L the number of lags, and N the number of points in the time series. The statistic G is defined as:

$$G = \frac{(RSS2 - RSS1)/L}{RSS1/(N-2L-1)} \quad (3)$$

In the case of Granger non-causality from X to Y, G follows a F distribution with degrees of freedom L and N-2L-1 [1]. The number of lags L in (1) and (2) can be defined assessing the quality of the statistical models. We adopted the AICc measure (AIC with a correction for finite sample sizes), and define L as the value minimizing the AICc in (2).

$$AICc = N \ln\left(\frac{RSS}{N}\right) + 2k + \frac{2k(k+1)}{N-k-1} \quad (4)$$

(Eq (4) assumes the errors are independent and normally distributed; k is the number of parameters in the model).

If L is fixed and equal to 1, equations (1) and (2) become:

$$Y_t = \alpha_0 + \alpha Y_{t-ly} + \beta X_{t-lx} + \epsilon_{1t} \quad (5)$$

$$Y_t = \alpha_0 + \alpha Y_{t-ly} + \epsilon_{2t} \quad (6)$$

We define ly and lx as the lags minimizing the RSS in equations (6) and (5) respectively. The experimental run was composed of 11 time series datasets of Yeast gene expression (2 time series from [2] and 9 time series from [3]). 2960 regulatory interactions involving the genes present in these datasets were retrieved from YEASTRACT [4]. For each time series, and for each pair of genes involved in a regulatory interaction, we estimated the statistic G for each direction of the interaction. For each time series, we assigned a p-value and a z-score to each G. The z-scores of the different time series (d=11) were combined following Stouffer's method:

$$Z \sim \frac{\sum_{i=1}^{d} Z_i}{\sqrt{d}} \quad (7)$$

The direction of each interaction was inferred as being the one with the highest z-score. In order to investigate the impact of the estimation of lags in the causal inference, we assessed the performance in three settings: in the setting M, multiple lags are considered (estimated using the AICc). In the setting E, only one lag is considered, and it is estimated (equations (5) and (6)). In the setting F, only the first lag is considered. In order to investigate the impact of taking into account auto-correlation, we only consider it in the AC settings (if not, then α_n=0 in equations (1) and (2), and the second degree of freedom of G becomes N-L-1).

RESULTS & DISCUSSION

Fig. 1 shows the proportion of correctly inferred interaction directions, considering only the 10%, 30% and 50% highest ranked interaction directions. The consideration of auto-correlation improves the inference performance, implying that this aspect is informative for the inference of causal interactions from time series. On the contrary, the use of multiple lags is not beneficial. The approach here described can be incorporated or used complementarily with current network inference methods for time series.

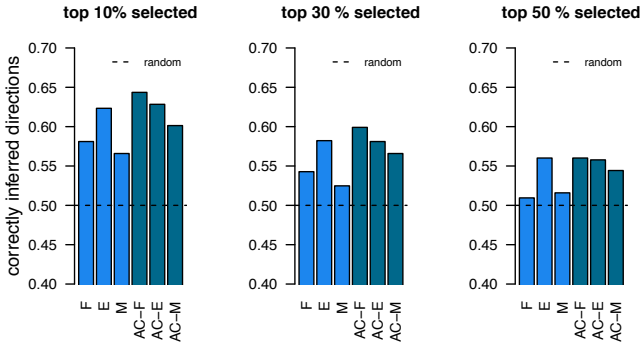


***FIGURE 1**. Proportion of correctly inferred directions, considering only the top ranked directions.*

REFERENCES

1. Brandt PT & Williams JT, Multiple Time Series Models, SAGE Publications (2007)
2. Simola DF *et al, Genome Biol.* **11** (2010)
3. Pramila T *et al. Genes Dev* **20(16)**, 2266-78 (2006)
4. Abdulrehman D *et al, Nucl. Acids Res.*, **39**, 136-140, (2011)

NETWORK INFERENCE BY INTEGRATING BICLUSTERING AND FEATURE SELECTION

Robrecht Cannoodt[1,2,3,4,], Joeri Ruyssinck[4], Katleen De Preter[3], Tom Dhaene[4], Yvan Saeys[1,2]*
*VIB Inflammation Research Center, Ghent, Belgium[1]; Department of Biomedical Molecular Biology, Center for Medical Genetics, Ghent, Belgium[2]; Department of Pediatrics and Genetics, Ghent University, Ghent, Belgium[3]; Department of Information Technology, Ghent University – iMinds, Ghent, Belgium[4]. *robrecht.cannoodt@ugent.be*

In order to develop better therapies to combat specific abnormalities present in the gene regulatory network (GRN) of cancer patients, it is crucial to gain a better understanding of regulatory networks in complex biological systems. An important class of methods in systems biology are network inference (NI) methods, which aim to reconstruct a GRN from high-throughput data (e.g. microarrays or next-generation sequencing).

INTRODUCTION

GENIE3[1] is a state-of-the-art method which employs feature selection to identify the best subset of regulators for each gene. While this method is amongst the best performing NI methods, it fails to take into account expected topological properties of a GRN: a GRN consists of modules, each of which consists of genes coregulated by a common set of regulators.

METHODS

We present BiGENIE, a method which takes the modular topology of a GRN into account. Figure 1 shows the difference in process between GENIE3 and BiGENIE. By firstly inferring modules – groups of genes coregulated by a common regulator – using several biclustering methods, the overall topology of the network is reconstructed. Subsequently, the regulator genes for each of the modules is inferred using GENIE3.

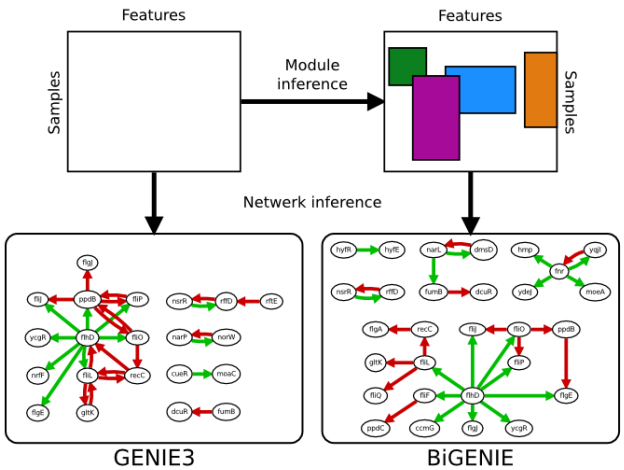


***FIGURE 1**. A comparison between GENIE3 and BiGENIE in terms of structure and the resulting inferred network. Green arrows represent correctly predicted interactions, while red arrows represent falsely predicted interactions.*

RESULTS & DISCUSSION

The Area-Under-ROC (AUROC) and Area-Under-Precision-Recall (AUPR) values are commonly used metrics to objectively evaluate the performance of NI methods. Table 1 contains AUROC and AUPR values of the GENIE3 and BiGENIE methods evaluated on 11 different datasets. The datasets consist of five small in silico networks from the DREAM4[2,3,4] competition, one large in vitro dataset from the DREAM5[5] competition, and five moderately sized networks generated by the GeneNetWeaver[6] tool (using the default settings).

Method	GNW200		DREAM4		DREAM5	
	AUROC	AUPR	AUROC	AUPR	AUROC	AUPR
BiGENIE	**0.762**	**0.171**	**0.771**	0.148	**0.642**	0.045
GENIE3	0.722	0.084	0.767	**0.187**	0.618	**0.094**
# genes	200		100		4511	
# samples	200		100		805	

***TABLE 1**. Evaluation of the BiGENIE and GENIE3 methods.*

From these results can be concluded that BiGENIE works very well on the GNW200 datasets. However, there is still room for improvement when BiGENIE is applied to networks of a different magnitude, as its AUPR values were less than GENIE3. By manually inspecting the results we noticed BiGENIE and GENIE3 do produce complementary results, so that these results could be combined to produce a better solution.

REFERENCES

1. Huynh-Thu V.A. *et al.* PLoS ONE **5**, e12776 (2010)
2. Marbach D. *et al.* Journal of Computational Biology **16**, 229–239 (2009).
3. Marbach D. *et al.* Proceedings of the National Academy of Sciences **107**, 6286–6291 (2010).
4. Prill R.J. *et al* PLoS ONE **5**, e9202 (2010).
5. Marbach D. *et al.* Nat Meth **9**, 796–804 (2012).
6. Schaffter T. *et al.* Bioinformatics **27**, 2263–2270 (2011).

NETTER: RE-RANKING GENE REGULATORY NETWORK PREDICTIONS USING STRUCTURE PROPERTIES

Joeri Ruyssinck[1,*], *Tom Dhaene*[1], *Yvan Saeys*[2,3].
Department of Information Technology (INTEC), Ghent University-iMinds, Ghent, Belgium[1];
VIB Inflammation Research Center, Ghent, Belgium[2];
Department of Respiratory Medicine, Ghent University[3],*Ghent, Belgium.* [*]*joeri.ruyssinck@intec.ugent.be*

Many algorithms have been proposed in the last decade that aim to deduce the gene regulatory network from high-throughput data such as microarray gene expression measurements. These algorithms typically produce a prioritized list of links between regulatory genes and their target genes based purely on connections in the data and fail to include typical structure properties of gene regulatory networks as prior knowledge. In this work we present Netter, a novel algorithm and flexible framework which uses a prioritized list of putative gene regulatory links as input and re-ranks this list based on various structure properties.

INTRODUCTION

The inference of large gene regulatory network (GRN) topologies from gene expression data is a challenging task, as the amount of genes in the network vastly exceeds the amount of available measurements. As a result, many network inference algorithms have been developed which use different strategies to overcome this inherent difficulty[1]. Surprisingly, almost none of the current state-of-art inference methods take into account general structure knowledge of a GRN. In contrast, although the topology of the GRN is very much dependant on the experimental conditions, general topological properties of GRNs have been described in literature[2].

The inclusion of complex and diverse topological information directly in the inference process of existing algorithms is non-trivial and offers little room for modifiability. Instead in this work, we propose and investigate a post-processing approach which we aim to be easily modifiable and extendable. The resulting algorithm, named Netter, uses as input any ranking of regulatory links sorted by decreasing confidence obtained by a network inference algorithm of choice. It then re-ranks the links based on structure properties, effectively penalizing regulatory links which are less likely to be true in the inferred network structure and boosting others.

METHODS

Netter works by defining an optimization problem in which we minimize a weighted sum of both desired structural properties of the predicted network and a regularization term penalizing divergence from the original prediction. This optimization problem is then solved several times using a simulated annealing approach, after which the obtained re-ranked lists are aggregated using average rank to obtain the final output ranking.

One of the structural properties that can be included in Netter is the desired graphlet[3] distribution of the network. Graphlets have been introduced as small connected non-isomorphic induced subgraphs of a larger network and differ from the concept of network motifs by the fact that an induced subgraph needs to contain all the edges between its nodes which are present in the parent network. Graphlets have been used in various other applications to characterize biological networks.

RESULTS & DISCUSSION

We have evaluated the performance of Netter on various artificially generated microarray datasets as well as on benchmark datasets available through the DREAM4 and DREAM5 network inference challenges. For this purpose, we selected two mutual information based network inference approaches, CLR and BC3NET, and two feature selection network inference approaches, GENIE3 and NIMEFI .

Our results show that by using only three suggested structural properties we can obtain a significant performance gain for all methods and all datasets. In particular, typically the top of the prediction ranking is significantly improved by Netter as is shown in Figure 1. Furthermore, our results indicate that the performance gain is robust with regard to various parameter settings and the weighing of the different included structural properties in the total fitness function. As such, we believe that Netter shows great potential as a post processing framework for gene regulatory network predictions.

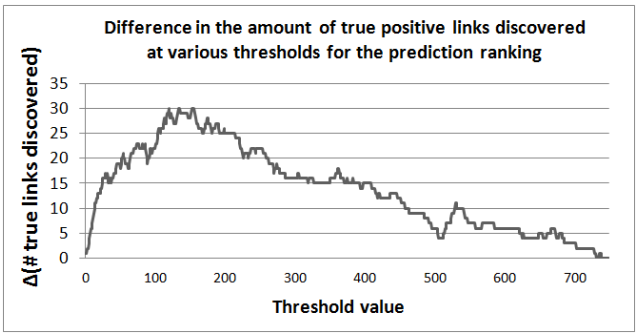


***FIGURE 1**. Comparison of the amount of true positive links discovered at various thresholds by the original prediction and the re-ranked prediction. The NIMEFI prediction of an artificially generated network is shown.*

REFERENCES

1. Marbach D. *et al.* Wisdom of crowds for robust gene network inference. *Nature Methods* **9,** 796–804 (2012).
2. Albert R. Scale-free networks in cell biology. *Journal of cell science* **118**, 4947-57 (2005).
3. Przulj N. *et al.* Modeling interactome: scale-free or geometric? *Bioinformatics* **20**, 3508–15 (2004).

THE RANK MINRELATION FOR TRANSCRIPTIONAL NETWORK INFERENCE

Patrick E. Meyer[1,*]
Machine Learning Group, Ib^2 , Université Libre de Bruxelles

Pairwise information measures such as correlations are heavily used in causal network inference algorithms mainly because they are fast and require few samples w.r.t multidimensional estimations. We propose a new set of asymmetric measures that aims at improving the detection of relevant variables. These new measures are called minrelation measures because of their similarity with correlation measures. Finally, we show through several experiments that these new measures are competitive with correlations in order to select relevant variables.

INTRODUCTION

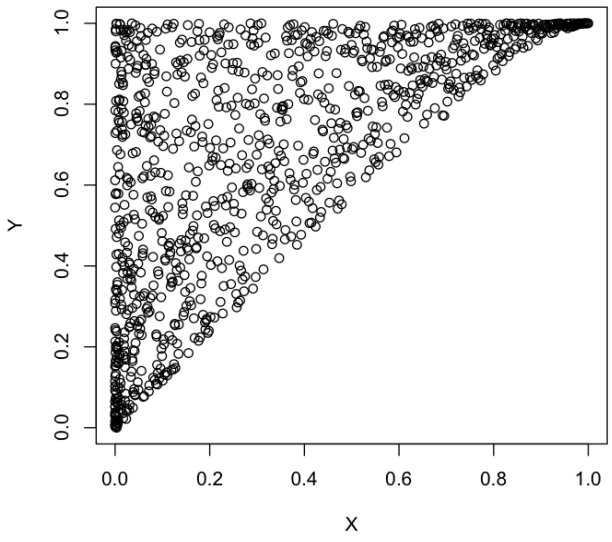
On the contrary to correlation, a minrelation is not symmetric. More explicitly, if a variable X exhibits a minrelation to Y then, as X increases, Y is likely to increases too, but, if X decreases, little can be said on Y values (except that the uncertainty on Y actually increases).

In this work, we define two different ways of measuring a minrelation. We also discuss the connection between minrelation and logical implication. Finally, we show through several experiments that these new measures are competitive with correlations in order to select relevant variables.

METHODS

Given two continuous random variables $X \in [0,1]$ and $Y \in [0,1]$, we define the measure of minrelation of X to Y as an estimate $p(X \leq Y)$. Respectively, the measure of majrelation of X to Y is the estimate of $p(X \geq Y)$.

For example (see Fig. 1), let m samples of $X \in [0,1]$ and $Y \in [0,1]$ be drawn such that $\forall (x_i, y_i), x_i \leq y_i, i \in \{1, ..., m\}$. In such case, there is a perfect minrelation between X and Y.



$X \to Y$	x_0	x_1
y_0	1	**0**
y_1	1	1

$p(X \to Y)$	x_0	x_1	$p(Y)$
y_0	0.33	0	0.33
y_1	0.33	0.33	0.66
$p(X)$	0.66	0.33	1

Table 1: Truth table of the implication and probability distribution of the probabilistic implication in the binomial case.

RESULTS & DISCUSSION

dataset	name	n	m	dataset	name	n	m
1	Ailerons	35	500	3	Triazines	58	186
2	Pol	26	500	4	Wisconsin	32	194

Table 5: Regression datasets, together with their number of variables n and number of samples m.

W/T/L of ι^*	vs $\rho^2(X,Y)$	vs $\rho^2(\overline{X},\overline{Y})$	vs $NMI(X,Y)$	vs $\hat{p}^*(X \to Y)$
LIN	**2/2/0**	**1/3/0**	1/2/1	**2/2/0**
RFOREST	2/0/2	**2/1/1**	2/0/2	**4/0/0**
SVM	**2/2/0**	1/2/1	1/2/1	**3/1/0**
Total	**6/4/2**	**4/6/2**	4/4/4	**9/3/0**

Table 6: Statistically significant wins/tails/losses, using 10-fold-cross-validated mean squared error returned by a linear regression, a random forest and a radial SVM averaged over subset sizes ranging from 2 to 10. For each of the four datasets, a ranking is returned using $\iota^*(X,Y)$ and compared to the ranking provided by each other criterion.

W/L of ι^*	vs $\rho^2(X,Y)$	vs $\rho^2(\overline{X},\overline{Y})$	vs $NMI(X,Y)$	vs $\hat{p}^*(X \to Y)$	#targets
KO1	**4/1**	**3/2**	1/4	**5/0**	5
KO2	**5/2**	2/5	**5/2**	**6/1**	7
KO3	**4/1**	**5/0**	**5/0**	**5/0**	5
KO4	**4/2**	3/3	**4/2**	**5/1**	6
KO5	3/6	4/5	4/5	**7/2**	9
MF1	0/5	**4/1**	**3/2**	**5/0**	5
MF2	**4/3**	**4/3**	**5/2**	**7/0**	7
MF3	**3/2**	**4/1**	**4/1**	**5/0**	5
MF4	3/3	3/3	**5/1**	**6/0**	6
MF5	**8/1**	**7/2**	**7/2**	**7/2**	9
Tot	**38/26**	**39/25**	**43/21**	**58/6**	64

Table 4: Wins/losses of $\iota^*(X,Y)$ vs other information measures in ranking strategies on target variables having more than 10 predictors in the 10 datasets of the DREAM4 competition. Column 1 indicates the dataset, column 2 indicates the number of variables having more than 10 predictors in that dataset and columns 3, 4, 5 and 6 reports the wins and losses of the two ranking methods on those target variables. A method wins if the average position of the predictors in the ranking is lower than for the other method. Bold notations are used when $\iota^*(X,Y)$ outperform the other criterion

THE CELL NUCLEUS HELPS REGULATE NOT ONLY THE DYNAMICS OF GENE EXPRESSION, BUT ALSO ITS NOISE

Jaroslav Albert[*] *& Marianne Rooman.*
Dept. of BioModeling, BioInformatics & BioProcesses, Université Libre de Bruxelles [*]*jalbert@ulb.ac..be*

One of the functions of the cell nucleus is to control molecular import and export, thus allowing for additional regulation of gene expression. Here we investigate, by way of the master equation and the Gillespie algorithm, what effects does segregation of biochemical processes into the nuclear and cytoplasmic compartments have on the intrinsic noise in protein concentration. For both, a single-gene case, and a four-gene coherent feed forward motif, we found that noise in the concentration levels in both the nucleus and cytoplasm are necessarily reduced. We found that in some cases a reduction in the standard deviation of up to 70% percent was achievable through parameter adjustment. The significance of this reduction, along with the necessity for accuracy in developmental gene circuits, suggest that it was partly to this end that the nucleus has evolved.

INTRODUCTION

The functioning of all biological systems is based on chemical reactions occurring at random. Although the dynamics of molecular concentrations in a population of cells or a tissue can be described with relatively high certainty, the situation in every individual cell is quite different: the molecular concentrations fluctuate in a manner that is unpredictable. Various mechanisms and devices exist that reduce these fluctuations[1]. One such device is the cell nucleus).

METHODS

To see how separation of processes which occur inside a cell nucleus from those taking place in the cytoplasm affects the intrinsic noise levels, we looked at the simplest case first: that of a single gene. The master equation for this system was solved exactly. As a first step, we considered solutions constrained to yield the average intra-nuclear protein concentration identical to the concentration of the same protein in a system with no nucleus. To do this, we defined a Euclidean-type distance measure that compared the dynamics of these two systems, and minimized it with respect to the system parameters, while keeping the fano factor at equilibrium below a certain threshold.

Next, we looked at a more complicated gene circuit: a four-gene coherent feedforward motif commonly found in developmental gene circuits of higher eukaryotes[2]. As in the simple case, we looked at the noise levels in protein concentrations without a nucleus, and compared them to the noise in the intra-nuclear protein concentrations of a system with a nucleus, all while keeping the average protein concentrations in the two systems equal. This latter case was studied using the Gillespie algorithm.

RESULTS & DISCUSSION

We found that the presence of a nucleus, and hence a separation of biochemical processes into compartments, necessarily reduces the noise levels both inside the nucleus and in the cytoplasm. The extent of this effect depends on the system parameters; however, out of the one-gene cases we studied, the best one achieved a reduction of 75% in the standard deviation of the protein concentration. For the motif case, we looked at two cases: one where the average dynamics in the two systems were (nearly) identical, and for which we only modified two of the system parameters common to both systems; and another where all system parameters were modified in order to reduce the noise, but which yielded average dynamics that were shifted in time and equilibrium concentration. However, this shift preserved the relative profile between the genes within the motif.

The quantity of interest was the distribution of threshold times – times at which the concentration of a particular protein reaches a certain threshold. For the two aforementioned cases, the reduction of uncertainty in the relative threshold times (standard deviation over the average) was 49% and 68% respectively.

While it may be argued that our model, as well as the system at hand, is too simple to confidently declare, "The cell nucleus has evolved in part as a necessity for highly accurate genetic circuitry to form," the enormity of the found effect on noise affords this proposition further investigation.

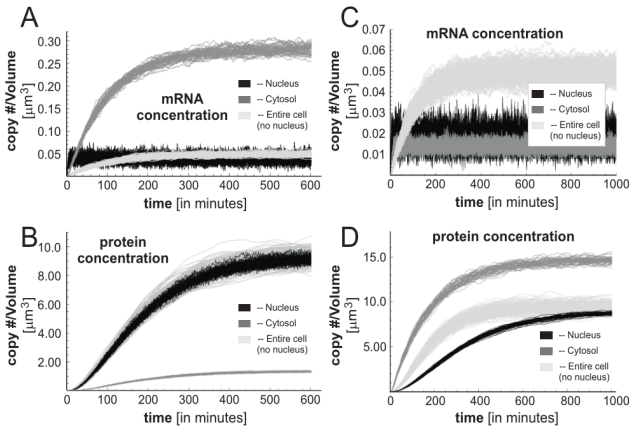


FIGURE 1*. Graphs showing simulations of mRNA and protein concentrations. The ensemble in light grey represents concentrations in a cell without a nucleus; the grey curves refer to cytoplasm, while those in black belong to the cell nucleus.*

REFERENCES

1. Sanchez A *et al*. *Ann Rev Biophys* **42**, 469-491 (2013).
2. Leon SB-TD & Davidson E Dev*elop. Biol.* **325** (12), 317-328 (2009).

DATA-DRIVEN VALIDATION OF GENE REGULATORY NETWORKS USING KNOCK-DOWN DATA

Catharina Olsen[1,*], *Gianluca Bontempi*[1], *John Quackenbush*[2] & *Benjamin Haibe-Kains*[3].
Machine Learning Group, Université Libre de Bruxelles, Brussels, Belgium[1]*; Computational Biology and Functional Genomics Laboratory, Dana-Farber Cancer Institute, Harvard School of Public Health, Boston, MA, USA*[2] *Ontario Cancer Institute, Princess Margaret Cancer Center, University Health Network, Toronto, Ontario, Canada*[3].
[*]*colsen@ulb.ac.be*

One strategy to better understand gene-gene interactions in biomedical problems is to infer gene regulatory networks from high throughput genomic data. However, the subsequent evaluation of these networks remains a difficult task because the number of known interactions is typically small and because experimental validation is expensive and time consuming. Recently, we proposed a purely data-driven validation procedure using knock-down experiments to quantitatively evaluate the quality of the inferred networks.

INTRODUCTION

We have come to accept that it is not single genes but networks of interacting genes that govern cellular processes[1]. Therefore when trying to understand underlying mechanisms that govern a disease such as cancer, it is not only crucial to infer gene regulatory networks but also to determine their quality[2]. Typically inferred networks are validated by i) comparing inferred interactions to known interactions stored in biological databases and ii) if a network has a lot of overlap with these known interactions, additional experiments are carried out to experimentally verify interactions. However, no quantitative validation strategy existed until our recent proposal to base the network validation on a number of single knock-down experiments[3]. This purely data-driven validation procedure then allows us to assess i) the quality of networks inferred using different algorithms and ii) the pertinence of known interactions retrieved from biological databases and published research to a specific biomedical problem.

METHODS

The basis of our validation strategy is a microarray data set in which was collected carrying out a number of single gene knock-down experiments. In the experiments we focused on eight genes related to the RAS pathway in two colorectal cancer cell-lines.

When knocking down one gene, a number of genes will be affected by this perturbation. We determine this set of affected genes for each of the eight knock-down experiments using the Wilcoxon rank sum test.

In an inferred gene regulatory network, these affected genes should be found somewhere in the knocked down gene's childhood. Therefore, after inferring a network from an independent data set, we can determine its quality in a three-step procedure as follows.

For each of the eight knock-downs, repeat:

- Determine the affected genes of the current knock-down
- Determine true positives, false positives and false negatives from the current knocked down gene's childhood.
- Compute performance measure for this part of the network.

RESULTS & DISCUSSION

For our experiments we infer networks from a large publicly available data set containing 292 human colorectal tumors[4]. We use different inference methods such as *GeneNet*[5] and *predictionet*[6]. The latter allows including known gene-gene interactions during the inference. We retrieve known interactions from biological databases and published research by using tools such as *Predictive Networks*[7].

Using the described validation procedure, we are able to show that:

- The networks inferred using *predictionet* are significantly associated with observed effects of new perturbations.
- The networks inferred using *predictionet* are more often significantly associated with the observed effects than those inferred using *GeneNet*.
- The known interactions are informative by themselves and when combining them with genomic data.
- There is only a small overlap in the information from genomic data and known interactions.

REFERENCES

1. Barabási AL & Oltvai ZN. *Nat Rev Genet* **5**, 101–113 (2004).
2. Fernald GH et al. *Bioinformatics* **27**, 1741–1748 (2011).
3. Olsen C et al. Inference of predictive gene networks from biomedical literature and gene expression data. Submitted to *Genomics* (2013).
4. http://expo.intgen.org/geo/
5. Opgen-Rhein R. & Strimmer K. *BMC Systems Biology* **1**, 37 (2007).
6. Haibe-Kains B et al. predictionet: Inference for predictive networks designed for (but not limited to) genomic data. *R package* (2012)
7. Haibe-Kains B et al. *Nucleic acids research*, 40, D866–D875 (2012).

CUTTER: GPU-BASED RECONSTRUCTION OF BIOLOGICAL NETWORKS FROM PERTURBATION EXPERIMENTS

Dragan Bosnacki[1,*], *Maximilian Odenbrett*[1], *Anton Wijs*[1], *Willem Ligtenberg* [1], *Peter Hilbers* [1].
Deps. of Bimedical Engineering[1], *Eindhoven University of Technology.* [*]*dragan@win.tue.nl*

We present CUTTER, a software tool for reconstruction of biological networks from perturbation experiments. In particular, CUTTER is focused on inference of transcription networks from knockout and knockdown experiments. The crucial idea is to apply transitive reduction for removing spurious edges using highly efficient parallel algorithms running on general purpose graphics processing units (GPUs). The empirical evaluation shows that CUTTER is more than hundred times faster than its sequential counterparts. Also, regarding the quality of reconstruction CUTTER outperforms the tools that have participated in the DREAM4 challenge.

INTRODUCTION

Perturbation techniques for reconstruction of genetic networks, like knockout or knockdown experiments, suffer from the problem of predicting direct interactions between genes that do not exist. Usually the cause of such falsely identified edges in the underlying graph of the network is the existence of an indirect interaction between the two genes. Transitive reduction of networks/graphs attempts to solve this problem by removing each direct interaction between two genes that are also connected via an indirect chain of interactions.

The existing algorithms for transitive reduction are sequential and might suffer from too long run-times for large genetic networks. They also exhibit the anomaly that some direct interactions are unnecessarily removed from the original network, like in the case of the feed-forward loops, which are common motifs in genetic networks.

METHODS

We present the tool CUTTER[1,4], based on a scalable parallel reduction algorithm using the quantitative aspects of the gene interactions. To this end we employ the concept of transitive reduction on weighted graphs. The edge weight corresponds to the (un)certainties of the interaction represented by the edge. The crucial idea behind this concept is to remove a direct interaction between genes only if there exists an indirect interaction which is strictly more certain than the direct one, according to the experimental data. This is a refinement of the removal condition for the unweighted graphs and avoids to a great extent erroneous removals of direct edges. Also we introduce the concept of transitive reduction with thresholds[1] which does not remove interactions which are above some certainty level. This reduces further the number of false negatives.

We develop efficient parallel algorithms for transitive reduction for general purpose graphics processing units (GPUs) for both standard (unweighted) and weighted graphs. the relative simplicity of our approach which can potentially be combined with other orthogonal reconstruction techniques in order to increase the performance. For instance, transitive reduction can be used as a pre- or post-processing stage of the existing algorithms.

RESULTS & DISCUSSION

Our experiments show that CUTTER for both weighted and unweighted graphs can achieve tremendous speed-ups (up to two orders of magnitude) compared to their sequential counterparts (Figure1).

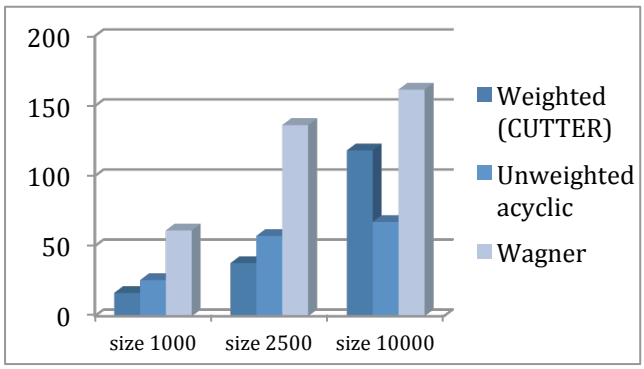


***FIGURE 1**. Speedups achieved by the GPU implementation of CUTTER with respect to sequential algorithms. On the x-axis is the size of the graphs in nodes and on the y-axis is the speedup. Weighted CUTTER is the sequential version of CUTTE; unweighted acylic is the version of the algorithm for acyclic graphs; Wagner is denoted the algorithm from [3].*

We tested the tool both on biological networks and using the DREAM4 benchmark InSiilico_Size100 sub-challenge [2]. In the recent years the DERAM challenges have became an unofficial standard for the evaluation of network reconstruction algorithms. With an overall score of **73,33** CUTTER is better than all submissions that participated in the sub-challenge [2,4].

One of the most interesting observations from the tests with the DREAM challenges is that taking into account the interaction sign (activation or inhibition) in the algorithms does not improve significantly the quality of the overall score.
This is an important result, since dealing with negative cycles greatly increases the complexity of the algorithms (they become NP-complete), while our algorithms are of polynomial complexity $O(n^3)$, where n is the number of genes in the network.

REFERENCES

1.D. Bosnacki *et al,. BMC Bioinformatics* **13,** 281-290 (2012). doi:10.1186/1471-2105-13-281
2. **The DREAM project** http://wiki.c2b2.columbia.edu/dream/index.php/The_DREAM_Project
3. A. Wagner, . *Bioinformatics* **17**(12)**:**1183-1197 (2001),.
4. CUTTER web page: http://www.win.tue.nl/emcmc/cutter

FASTCORE: AN ALGORITHM FOR FAST RECONSTRUCTION OF CONTEXT-SPECIFIC METABOLIC NETWORK MODELS

Nikos Vlassis[1,*], Maria Pires Pacheco[2] & Thomas Sauter[2].
Luxembourg Centre for Systems Biomedicine[1] & Life Sciences Research Unit[2], University of Luxembourg.
[*]nikos.vlassis@uni.lu

Cell metabolism is known to differ significantly from cell to cell and over different contexts. Hence, the efficient reconstruction of context-specific metabolic models is of high interest. We describe FASTCORE, a fast algorithm for the reconstruction of compact context-specific metabolic network models. The algorithm takes as input a global metabolic model (such as Recon 2), and a 'core' set of reactions that are known to be active in a given context. Then it produces a context-specific network model that contains the core set and a minimal set of extra reactions. FASTCORE is order of magnitudes faster than other algorithms, typically obtaining a genome-wide reconstruction in a few seconds.

INTRODUCTION

To maximize the predictive power of a metabolic network model when conditioning on a specific context (e.g., liver), recent efforts go into the development of *context-specific* metabolic network models[1,2]. These are models that are derived from global models such as Recon 2, but they contain (ideally) only those reactions that are expected to be active in the given context. The computational problem is: Given a 'core' set of reactions C and a global network N, find a *(flux-) consistent* subnetwork of N that contains all reactions from C and a minimal set of additional reactions.

METHODS

It is easy to show that a consistent induced subnetwork of the global network can be defined via a set of *modes* of the latter:

Theorem 1. *Let V be a set of modes of the global network N, and let A be the union of the supports of these modes. The subnetwork induced by A is consistent.*

Hence, if S_N denotes the stoichiometric matrix of N, the optimal reconstruction problem can be expressed as:

$$\begin{aligned} \min_{\mathcal{V}} \quad & card(\mathcal{A}) \\ \text{s.t.} \quad & \mathcal{A} = \bigcup_{v \in \mathcal{V}} supp(v) \\ & \mathcal{C} \subseteq \mathcal{A} \\ & \forall v \in \mathcal{V}: \quad S_{\mathcal{N}} v = 0, \; v \in \mathcal{B}. \end{aligned}$$

The last equation defines the modes of N, i.e., it enforces mass-balance at steady state (B defines the flux bounds).

FASTCORE is a *greedy* search strategy for computing the set of modes V in the above program, reminiscent of greedy heuristics for the related *set covering* problem[3]. In each iteration FASTCORE adds to the set V a new mode of N, constrained to have *sparse* support outside C. The latter is achieved by an L_1-norm regularized linear program, which provides a convex relaxation to the minimum cardinality constraint[4]. We refer to the full version of our paper for more technical details[5].

RESULTS & DISCUSSION

We used the FASTCORE algorithm to reconstruct a liver-specific metabolic network model from Recon 1 ($|N|$ = 2469 after removing blocked reactions), and we compared against the MBA algorithm[2]. We applied the two algorithms in two settings. The first setting involves a 'standard' liver-specific input reaction set[2] that is based on 779 'high' core and 290 'medium' core reactions (the latter set supported by weaker biological evidence than the former). To allow a comparison with FASTCORE, we defined a single core set as the union of the high and medium core reaction sets, and we applied the two algorithms on this core set. The second setting involves a 'strict' reaction set[2] that contains 1083 high core reactions and no medium core reactions, and therefore allows a direct comparison with FASTCORE.

The results are summarized in Table 1. In both settings, FASTCORE is several orders of magnitude faster than MBA, achieving a full reconstruction of a liver specific model in about one second, using a much smaller number of LPs. The reconstructed models by FASTCORE are also more compact than those obtained by MBA, with a difference of 70-80 non-core reactions. The two algorithms turned out to use alternative transporters to connect the core reactions: In the standard liver model, 46 out of 59 reactions that are present exclusively in the FASTCORE reconstruction are transporter reactions or other reactions that are not associated to a specific gene (and thus are not sufficiently supported in the core set), whereas in MBA the corresponding numbers are 116 out of 139 reactions. It is important to emphasize that the difference in the number of added non-core reactions between MBA and FASTCORE is the result of the different optimization approaches taken by the two algorithms, and hence no biological relevance should be attributed to each reconstruction other than the one implied by the makeup of the core set. From this point of view, FASTCORE performs in general better than MBA, as it tends to add fewer unnecessary reactions. In the full version of the paper we provide additional results[5].

	liver core set ($\#\mathcal{C} = 1069$)				strict liver core set ($\#\mathcal{C} = 1083$)			
	$\#\mathcal{A}$	IR*	#LPs	time‡	$\#\mathcal{A}$	IR	#LPs	time
MBA	1826	1573	72279	7383	1888	1630	71546	6730
FASTCORE	1746	1546	20	1	1818	1627	20	1

*number of intracellular reactions
‡the reported time (in seconds), as well as the number of LPs, refer to a single pruning step of MBA, whereas $\#\mathcal{A}$ and IR refer to the full MBA.

***TABLE 1**. Comparing FASTCORE to MBA[2] on two liver-specific reconstruction problems.*

REFERENCES

1. Becker S, Palsson BO. *PLoS Comp Biol* 4: e1000082 (2008).
2. Jerby L et al. *Mol Syst Bio* 6:401 (2010).
3. Chvátal V. *Math Oper Res* 4: 233-235 (1979).
4. Julius AA et al. *IEEE Conf. on Decision and Control*. 762-767 (2008).
5. Vlassis N, et al. *ArXiv*:1304.7992 (2013).

NETWORK DEREGULATION ANALYSIS IN COMPLEX DISEASES VIA THE PAIRWISE ELASTIC NET

Nikos Vlassis[*] *& Enrico Glaab.*
Luxembourg Centre for Systems Biomedicine, University of Luxembourg. [*]*nikos.vlassis@uni.lu*

Complex diseases like neurodegenerative or cancer disorders are characterized by deregulations in multiple genes and proteins. Previous research has shown that neighboring genes in a molecular network tend to undergo coordinated expression changes. We describe an approach that allows identifying such jointly differentially expressed genes from input expression data and a graph encoding pairwise functional associations between genes (such as protein interactions). We cast this as a feature selection problem in penalized two-class (cases vs. controls) classification, and we propose a novel Pairwise Elastic Net penalty that favors the selection of discriminative genes according to their connectedness in the interaction graph. Experiments on microarray gene expression data for Parkinson's disease demonstrate marked improvements in feature grouping over competitive methods.

INTRODUCTION

We assume supervised *gene expression* data $\{\mathbf{x}_i, y_i\}_{i=1,\ldots,n}$, with features $\mathbf{x}_i \in R^p$ (with $p>>n$) and class labels $y_i = \pm 1$, and a *graph* that encodes pairwise functional associations between genes (e.g., a protein-protein interaction network). We are interested in computing a *sparse* set of discriminative genes that form a large *connected* subgraph of the input graph.

METHODS

We adopt a model-based approach using penalized logistic regression. This involves finding weights $\mathbf{w} \in R^p$ and $\nu \in R$ that solve the program:

$$\min_{\mathbf{w},\nu} \; f(\mathbf{w},\nu) + \lambda\,\Omega(\mathbf{w}),$$

where $f(\mathbf{w},\nu)$ is the standard expected logistic loss, and $\Omega(\mathbf{w})$ is a penalty term that regularizes $\mathbf{w}$, with tradeoff parameter $\lambda>0$. We propose an novel penalty function $\Omega(\mathbf{w})$ that penalizes the differences between the *absolute* values of weights of neighboring features in the graph:

$$\Omega(\mathbf{w}) = \sum_{i=1}^{p} \left[\sum_{j=1}^{p} A_{ij}|w_i| - \sum_{j=1}^{p} A_{ij}|w_j| \right]^2 + 2\Delta\|\mathbf{w}\|_1^2 ,$$

where A is the *adjacency* matrix of the input graph and Δ its maximum *degree*, respectively, and $\|\mathbf{w}\|_1$ is the L_1 norm of the weight vector. Our main technical result is the following:

Theorem 1. *The penalty function $\Omega(\mathbf{w})$ is convex in $\mathbf{w}$.*

The proof involves showing that $\Omega(\mathbf{w})$ is an instance of the *Pairwise Elastic Net* (PEN)[2] family of penalties. The above gives rise to convex program that can be efficiently solved with TFOCS[4] (a few seconds for a typical run with n=100, p=1000). At optimality, feature i is selected if $\mathbf{w}_i$ is nonzero.

RESULTS & DISCUSSION

We compared our approach to penalized logistic regression using the *Lasso*[1], the *Elastic Net*[1], and the PEN[2,3] penalties. We used public microarray gene expression data from a case-control study of *Parkinson's disease*, with nearly balanced numbers of cellular samples for patients and unaffected controls[5]. As input feature graph we used a genome-scale protein-protein interaction network assembled from multiple public databases. As seen in Figure 1, our approach extracts significantly more connected feature groups than the other methods, across the whole regularization path: In a typical run, 34 out of 51 genes (67%) selected by our method formed a connected subgraph of the input graph, vs. about 30% for a related graph-based PEN penalty[3], and much less for the other methods. Subsequent text-mining analysis revealed that several of the genes in this subgraph have functional annotations that are strongly associated with Parkinson's disease (details omitted). Our approach exhibited similar cross-validation performance than all other tested methods, and hence the feature grouping is not obtained at the expense of lower predictive power. The resulting deregulated gene clusters enable a network-level interpretation of molecular changes in the disease, and they can serve as candidates for combinatorial biomarkers to overcome the lack of robustness often observed for single-gene markers.

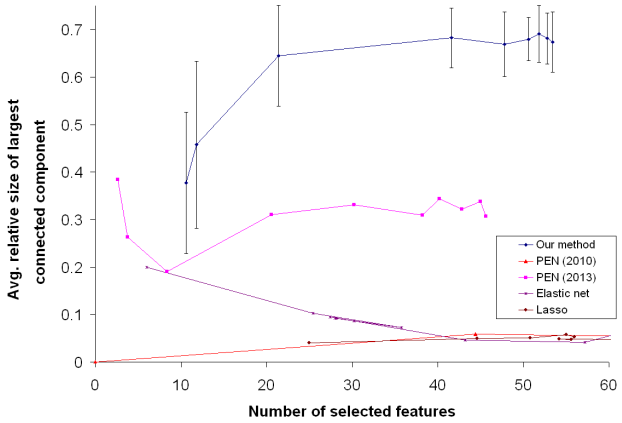


***FIGURE 1**. Feature grouping obtained by different methods.*

REFERENCES

1. Hastie T et al. *The Elements of Statistical Learning,* Springer, 2nd ed. (2009).
2. Lorbert A, et al. *Int. Conf. on Artificial Intelligence and Statistics (AISTATS)* (2010).
3. Lorbert A & Ramadge PJ. *Int. Conf. on Acoustics, Speech, and Signal Processing (ICASSP)* (2013).
4. Becker SR et al. *Mathematical Programming Computation* **3**:165–218 (2011).
5. Zhang Y et al.. *American Journal of Medical Genetics Part B*, **137**:5–16 (2005).

NoDe and CATCh: Two Algorithms to Get More Accurate 16S rRNA Sequencing Data.

Mohamed Mysara [1-3], Yvan Saeys [4], Natalie Leys[1], Jeroen Raes [2,3], Pieter Monsieurs [1*].
Unit of Microbiology, Belgian Nuclear Research Centre SCK•CEN, Mol; Belgium[1;] Department of Bioscience Engineering, Vrije Universiteit Brussel VUB, Belgium[2]; Department of Structural Biology, Vlaams Instituut voor Biotechnologie VIB, Belgium[3]; Data Mining and Modeling group, VIB Inflammation Research Center, Belgium[4]. *pmonsieu@sckcen.be

Next generation sequencing is has created a wide range of new applications, also in the field of microbial diversity, Yet when used in 16S rRNA biodiversity studies, it suffers from two important problems: the presence of PCR artefacts (called chimera) and sequencing errors. In this work two artificial intelligence-based algorithms are proposed, CATCh and NoDe to handle these two problems. A benchmarking study was performed comparing CATCh and NoDe with other state-of-the art tools, showing a clear improvement in chimera detection and reduction of sequencing errors respectively.

Introduction

The revolution in new sequencing technologies has led to an explosion of possible applications, including microbial biodiversity studies in the environment by bacterial 16S rDNA sequencing. However all sequencing technologies suffers from the presence of erroneous sequences, i.e. (i) chimera, introduced by wrong target amplification in PCR, and (ii) sequencing errors originating from different factors during the sequencing process. As such, there is a need for effective algorithms to remove those erroneous sequences to be able to accurately assess the microbial diversity.

Methods

First, a new algorithm called CATCh (Combining Algorithms to Track Chimeras) was developed integrating the output of existing chimera detection tools into a new more powerful method. Second, NoDe (Noise Detector) was introduced, an algorithm that identifies those positions in 454-sequencing reads likely to contain an error, and subsequently clusters those error-prone sequences with correct reads resulting in error-free consensus reads. This leads to a decrease in the number of reads or nucleotides that is disregarded by the current state-of-the-art denoising algorithms. Third, NoDe and CATCh were combined with a straight-forward pre-processing approach, creating a 454 16S rRNA sequencing analysis pipeline. Our algorithms, NoDe and CATCh, are freely available, and can easily be integrated with other 16S rRNA data analysis platforms (e.g. Mothur [3]).

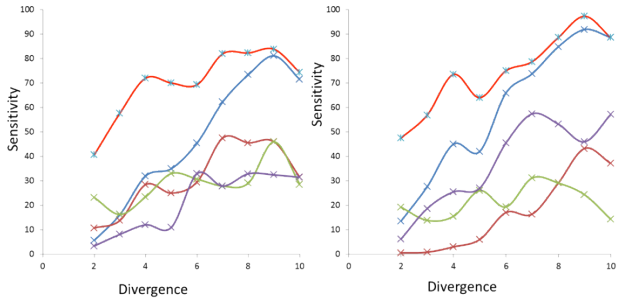


Figure 2. *Plot indicating the effect of applying 5% indels (shown on the left) and 5% mismatches (shown on the right), mimicking the biological variations, on the performance of different tools. The newly developed algorithm CATCh was found to outperform other existing tools.*

Results & Discussion

Via a comparative study with other chimera detection tools, CATCh was shown to outperform all other tools, thereby increasing the sensitivity with up to 14% (see Figure 1). Similarly, NoDe was benchmarked against state-of-the-art denoising algorithms, thereby showing significant improvement in reduction of the error rate (reduction of 15 to 55%) (see Figure 2), combined with an minimal rejection of sequencing data retained after cleaning (20% more) and decrease in computational costs (90% faster). The cut-off parameters needed for the analysis of 454 pyrosequencing data were optimized, based on a Mock community (i.e. a known mixture of 18 bacterial species).

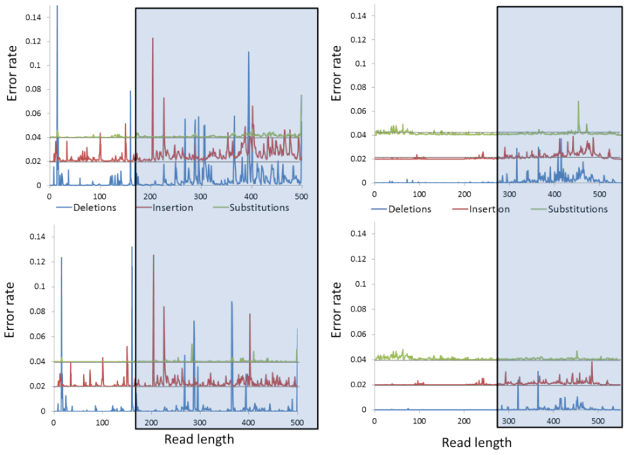


Figure 2. *Plot showing different error types (substitutions, deletions, mismatches) along the position in the read. The error rate (y-axes) increase with the increasing length (x-axe). After applying the new NoDe denoising algorithm, a significant reduction in error rate is observed (highlighted above) using two different test data sets, A[1] and B[2].*

Conclusively, introducing CATCh and NoDe into an existing 454 pre-processing pipeline, increases the overall reliability of the data and corrects existing sequencing errors. Further work implies fine-tuning toward other platforms (e.g. Illumina MiSeq).

References

1. Gilles A *et al. BMC genomics* **12**, 245 (2011).
2. Schloss PD *et al. PloS one* **6**,e27310 (2011).
3. Schloss PD et al. *Appl. Environ*. Microbiol. **75**:7537–41 (2009).

APPLICATIONS OF LARGE-SCALE GENOME-WIDE DNA METHYLATION PROFILING

*Tim De Meyer**, Geert Trooskens, Klaas Mensaert, Sandra Steyaert, Matthijs Vynck, Simon Denil & Wim Van Criekinge*

Department of Mathematical Modelling, Statistics and Bioinformatics & Nucleotides 2 Networks, Ghent University

**Tim.DeMeyer@UGent.be*

DNA methylation is an important epigenetic modulator of gene expression in health and disease. Genome-wide DNA methylation profiling can be performed using a combination of methylation-specific affinity purification and subsequent massive parallel sequencing (MethylCap-seq). Whereas this methodology is typically used for biomarker discovery, locus unbiased large-scale DNA methylation profiling has far more applications, e.g. to infer pathways and functional regions relevant to methylation associated differentiation, the identification of mono-allelic methylation and the detection of repeat and virus methylation.

INTRODUCTION

DNA methylation is an important epigenetic modulator of gene expression in health and disease. Most studies in the field focus on the identification of biomarkers, typically using methods that only assess (large sets of) preselected loci. There are however several methods, such as MethylCap-seq (aka MBD-seq), that allow for (virtually) genome-wide analyses beyond biomarker identification.[1] Here we demonstrate the usefulness of the cost-affordable MethylCap-seq methodology for research in several related application domains: (1) DNA methylation is relevant for cellular differentiation, but it is unclear which functional regions of the DNA methylome are responsible; (2) Mono-allelic gene-expression is involved in several non-Mendelian inherited genetic disorders, and often regulated by DNA methylation, but only few mono-allelically methylated loci have been identified; (3) Expression of genomic repeat regions and viral sequences can be silenced by DNA methylation, and might therefore also be picked up by MethylCap-seq, but the practical usefulness remains to be demonstrated.

METHODS

Figure 1 summarizes the MethylCap-seq methodology, targeting 5-methylcytosine as epigenetic modification.

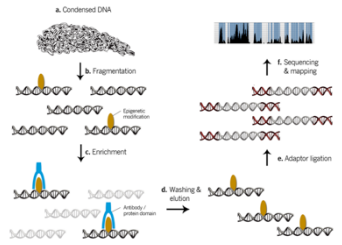


FIGURE 1. *MethylCap-seq (using Methyl Binding Domain – MBD – for capture)*

Mapping of sequenced reads was performed using BOWTIE[2] and data-analytical pipelines consisted of a combination of PERL and R CRAN scripts. For biomarker discovery, data was summarized using the Map of the Human Methylome (www.biobix.be/mhm), and a Poisson model was used to identify ignificant methylation. Loci methylated in tumor but not control cell lines and featuring higher expression upon demethylation were validated in an independent population using Methylation Specific PCR.

For the tissue differentiation study, 10 tissues x 5 samples (mainly tumor) were included. Mono-allelic methylation identification was assisted by SNPs present in MethylCap-seq reads that support deviation from Hardy-Weinberg equilibrium. The European Nucleotide Archive (ENA)[3] and RepBase[4] databases were used to identify resp. viral and repeat sequences in the MethylCap-seq reads.

RESULTS & DISCUSSION

An example of biomarker identification was the generation of a panel of 4 prognostic methylation markers with cumulative effects for clear cell renal cell carcinoma (collaboration with Prof. Dr. M. Van Engeland, results from independent validation study, Figure 2).

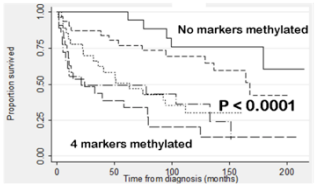
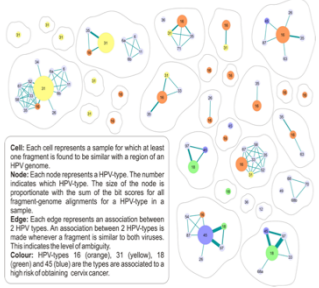


FIGURE 2. *Biomarker study*

For the identification of loci involved in tissue differentiation, all loci were ranked based on variance of (log) methylation degrees and enrichment analysis was performed on the top 10k variables. Not unexpectedly this revealed enrichment in promoter regions, but also in exon regions, and in several pathways (e.g. GPCRs, focal adhesion), indicating relevance of these pathways and exon methylation in differentiation.

At an FDR of 0.1, 80 loci (43 genes) could be identified that were featured by mono-allelic methylation, including the known imprinted IGF2/H19 locus. Importantly, the method can also be used for ChIP-seq approaches targeting e.g. histone modifications.



A pipeline to map "non-human" MethylCap-seq reads to RepBase and ENA was applied on a set of 29 cervical cancer samples, identifying the cause of this tumor, HPV,[5] in 22 cases (Figure 3)

FIGURE 3. *Identification of HPV in cervical cancer*

REFERENCES

1. Mensaert K *et al*. *EMM* (invited review, under peer review).
2. Langmead B & Salzbert S. *Nature Methods* **9**, 357-359 (2012).
3. Leinonen R *et al*. *Nucleic Acids Res* **39**, D19-D21 (2011).
4. Jurka J *et al*. *Cytogenetic Genome Res* **110**, 462-467 (2005).
5. Walboomers JMM *et al*. *J Pathol* **189**, 12-19 (1999).

SNP-GUIDED IDENTIFICATION OF MONOALLELIC DNA-METHYLATION EVENTS FROM ENRICHMENT-BASED SEQUENCING DATA

Sandra Steyaert[1,*], *Geert Trooskens*[1], *Ayla De Paepe*[1], *Simon Denil*[1], *Klaas Menschaert*[1], *Wim Van Criekinge*[1] *& Tim De Meyer*[1].

Department of Mathematical Modelling, Statistics and Bioinformatics, Ghent University, Ghent, Belgium[1].

[*]*Sandra.Steyaert@Ugent.be*

Monoallelic DNA-methylation is a central player in the regulation of monoallelic gene expression, of which dysregulation has been linked to certain non-Mendelian inherited disorders. Combining the SNP- and methylome profiles, obtained with MBD-seq, we present a new methodology to identify monoallelically methylated genomic regions. This methodology consists of SNP-identification, sequencing error filtering, and Hardy-Weinberg theorem based detection of monoallelic methylation. When applied on a diverse set of 344 human samples, this approach resulted in the identification of 80 regions with significant monoallelic methylation of which 25 have already been linked to imprinting. Furthermore, significant enrichment of promoter methylation was found for these loci, further corroborating the outcome of our method. Importantly, the developed approach can be easily adjusted to use for other enrichment-based sequencing technologies, like for example ChIP-seq.

INTRODUCTION

Monoallelic gene expression is typically initiated early in the development of an organism and stably maintained. Erroneous monoallelic expression has already been linked to some genetic disorders. DNA-methylation plays a significant role in the regulation of monoallelic expression. Whereas MBD-seq and ChIP-seq are widely used genome-wide methods to detect DNA methylation, they do not provide information regarding unmethylated alleles. Here, we present a data-analytical methodology that circumvents this problem by means of the Hardy-Weinberg theorem and thereby enables screening for genes that exhibit monoallelic DNA-methylation and thus might regulate monoallelic expression.

METHODS

Methyl-CpG binding domain based sequencing (MBD-seq), which combines enrichment of methylated DNA-fragments by methyl-binding domain based affinity purification with massively parallel sequencing (Illumina GAIIx & HiSeq, paired end), was used to profile the DNA-methylation pattern of 334 diverse human samples.

Starting from this MBD-seq data and making use of the public NCBI SNP-archive (dbSNP), the obtained non-duplicate, uniquely mappable sequence reads (Bowtie) were screened for SNPs. After the determination of the base frequency per sample, only those SNP-loci with an adequately coverage and allele frequency were retained. In order to further reduce the effect of sequencing errors, an additional filtering step was performed by comparing the chance of a sequencing error with the chance of detecting genuine SNPs. For each single SNP-locus, the Hardy-Weinberg theorem can be applied to evaluate whether the observed frequency of samples featured by biallelic methylation is lower than randomly expected. If for a sample both alleles are observed at a SNP-locus, this sample is considered heterozygous at this locus. If only one allele is observed at this locus, the sample is considered homozygous. Taken together for all the samples, the identification of SNP-loci with monoallelic methylation results in finding a significant discrepancy between the observed and the theoretical heterozygous fraction. Indeed, in case of perfect monoallelic methylation, the observed heterozygous fraction equals 0. Using a permutation approach, loci with a p-value smaller than a selected FDR can be assumed to be significantly monoallelically methylated.

RESULTS & DISCUSSION

Figure 1 summarizes the results of the pipeline for all (diploid) chromosomes. The inner circle shows the histograms of all SNPs found in a specific region, whereas the outer circle depicts the histograms of the significant SNPs in that same region, normalized to the amount of SNPs found in that region. With an FDR of 0.1, we identified 80 significant SNPs.

Of these 80 loci, 49 were located in genic regions of which 25 have already been linked to imprinting, for example the H19/IGF2 locus. Interestingly, because a very diverse set of samples was used in the analysis, the found loci can assumed to be overall imprinted in somatic tissue. Not unexpectedly, an enrichment analysis demonstrated enrichment for monoallelic methylation in promoter regions (p=0.002), further corroborating the outcome of this study.

Importantly, next to MBD-seq, our approach also opens the door to other enrichment-based sequencing applications. In a likewise manner as described above, the developed methodology can also be used on ChIP-seq data in order to detect possible monoallelic protein-DNA binding events.

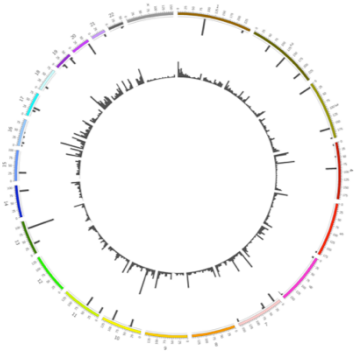


***FIGURE 1**. Circular representation of the genomic distribution of the 80 significant loci. Chromosomes are divided in regions of 5,000,000 bp.*

MINING THE GARBAGE FRAGMENTS OF METHYLATION-SPECIFIC ENRICHED DNA SEQUENCING

Klaas Mensaert[1], *Geert Trooskens*[1], *Simon Denil*[1], *Ed Schuuring*[2], *Bea Wisman*[2], *Wim Van Criekinge*[1], *Olivier Thas*[1] *& Tim De Meyer*[1,*].

Dept. of Mathematical Modeling, Statistics and Bioinformatics & Nucleotides 2 Networks, Ghent University[1]*; Dept. of Pathology and Gynecological Oncology, University of Groningen, University Medical Center Groningen*[2].

[*]*Tim.DeMeyer@UGent.be*

The combination of methylation-specific affinity purification and NGS sequencing (MethylCap-seq) is a cost-efficient technique for the genome-wide profiling of DNA-methylation. In order to identify methylated regions, sequenced fragments are unambiguously mapped to a reference genome. However, for a typical experiment, a large portion of the fragments cannot be stringently mapped. These fragments include repeats and viral sequences. We here present a pipeline in order to mine these fragments for additional information. This pipeline was successfully validated in a cervical tumor study.

INTRODUCTION

DNA-methylation, an important epigenetic mark, is associated with transcriptional regulation and alternative splicing. Its clinical importance can be observed in various diseases such as cancer. By combining methyl binding domain based affinity purification and sequencing, MethylCap-seq (aka MDB-seq) allows for a putatively genome-wide overview of the methylome[1]. An outline of this cost-efficient method is depicted in Figure 1. As regions are considered to be methylated when fragments are mapped onto them, sequenced fragments should be unambiguously mapped to a reference genome. However, for MethylCap-seq experiments in human, about 35% of the fragments cannot be stringently mapped onto the reference genome. These fragments potentially originate from repeats and viral genomes, which might be involved in human pathologies.

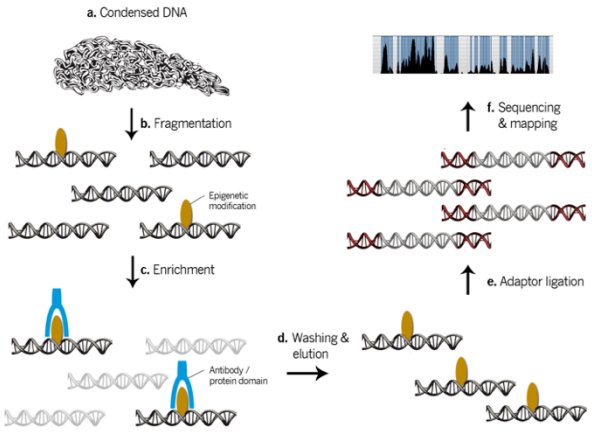


FIGURE 1. *Outline of MethylCap-seq* [2]

METHODS

A pipeline was constructed in order to mine fragments that could not be uniquely mapped to the human reference genome. Bowtie[3] was used for mapping with tolerant parameters, giving more certainty that unmapped reads are not of human origin. Viral particles and repeats were subsequently identified by mapping against respectively the viral genome sequences of the European Nucleotide Archive[4] and the Repbase database[5]

RESULTS & DISCUSSION

Validation of the viral part of the pipeline was performed in a study on cervical cancer, a tumor type known to be virtually always caused by human papillomavirus (HPV)[6]. Indeed, HPV was found in most of the samples originating from carcinogenic tissue, although as well in several normal tissue samples. The HPV-types that were found do correspond well with those that are highly risk increasing for cervical cancer[5]. Experimental validation of these HPV-type occurrences, including in the normal samples, is currently being performed. A positive control, the retrovirus K113 was identified in all samples[7]. Repeat methylation quantification can be a useful tool to discriminate between malignant and control tissues.

REFERENCES

1. De Meyer T *et al. Plos One* **8**, 1932-6203 (2013).
2. Mensaert K *et al. EMM* (under peer review).
3. Langmead B *et al. Genome Biol* **10**, R25 (2009).
4. Leinonen R *et al. Nucleic Acids Res* **39**, D19-D21 (2011).
5. Jurka J *et al. Curr Opin Struct Biol* **8**,333-337 (1998).
6. Clifford GM *et al. Br J Cancer* **89**, 101 -105 (2003).
7. Jha A. *et al. Mol Biol Evol* **26**, 2617-26 (2009).

ASSESSING THE OUTCOME OF 16S RDNA-BASED COMMUNITY ANALYSIS BY COMPREHENSIVE SIMULATIONS

Ali May[1, 2,*], *Sanne Abeln*[2], *Bernd W. Brandt*[1, *]

Department of Preventive Dentistry, Academic Centre for Dentistry Amsterdam (ACTA), University of Amsterdam and VU University Amsterdam[1]*; Centre for Integrative Bioinformatics VU (IBIVU), VU University Amsterdam*[2]*. *a.may@vu.nl, b.brandt@acta.nl*

The analysis of 16S rDNA pyrosequencing data requires a repertoire of computational methods for the pre-treatment and clustering of sequences, the evaluation of which remains confined to *in vitro* studies where mock samples are typically much less diverse than environmental samples. We generated complex *in silico* datasets with representative properties and evaluated two widely used pre-processing methods in combination with seven clustering algorithms. Analyses of moderately complex datasets (200 OTUs) showed that diversity overestimation did not occur and that there was even a slight underestimation. We identified Esprit-Tree and TBC as two algorithms that yielded the most accurate clusters when no denoising was applied on the data prior to clustering, while after denoising no significant difference was observed between algorithm performances.

INTRODUCTION

16S rDNA pyrosequencing is a powerful approach for the recovery of microbial compositions. Data pre-processing and clustering methods used for the analysis of 16S rDNA pyrosequencing studies have been limited to *in vitro* studies where mock samples are typically orders of magnitude less diverse than environmental samples. Here, we first analyse two low-complexity datasets to derive several data characteristics, which are then used to generate complex datasets. We apply two widely used data pre-processing tools in combination with seven different clustering algorithms to these datasets and evaluate the accuracy of the resulting clusters using the normalized mutual information (NMI) and the number of OTUs formed.

METHODS

Two GS-FLX Titanium datasets that were derived from pyrosequencing 16S rDNA (V5-V7 HVR) from a mock community of 15 oral bacteria were analyzed. Data features including the read length distribution, the abundance of chimeric sequences and the rate of errors introduced by PCR and sequencing were calculated. Using these properties, we first replicated one of the *in vitro* mock datasets *in silico* to establish that real data properties can be approximated in simulations. Next, we simulated 50 complex datasets of 40,000 reads by randomly selecting 250 reference sequences from the CORE[1] oral microbiome database. These datasets were separately pre-processed in five different ways: no cleaning (NC), chimera checked (CC), denoised (D), denoised and chimera checked (DCC) and chimera checked and denoised (CCD). Seven different clustering algorithms (CD-HIT, DNACLUST, Esprit-Tree, UCLUST, USEARCH, CLUSTOM and Taxonomy-Based Clustering (TBC)) were evaluated after each pre-processing approach by calculating the NMI score and the number of OTUs.

RESULTS & DISCUSSION

Primary characteristics of *in vitro* pyrosequencing data were successfully replicated in simulations. The difference between the clustering algorithms in terms of NMI scores and the numbers of OTUs became less pronounced when they were applied after pre-processing approaches that involved denoising. At species level, the OTUs formed by clustering the non-erroneous (reference) datasets expectedly yielded higher NMI scores than the OTUs formed by clustering the pre-processed datasets. At genus level, however, denoising followed by chimera checking, as well as chimera checking followed by denoising resulted in NMI scores higher than reference clusterings. Clustering by Esprit-Tree and TBC resulted in consistently high species and genus-level NMI scores, as well as numbers of OTUs close to those formed by clustering the reference dataset.

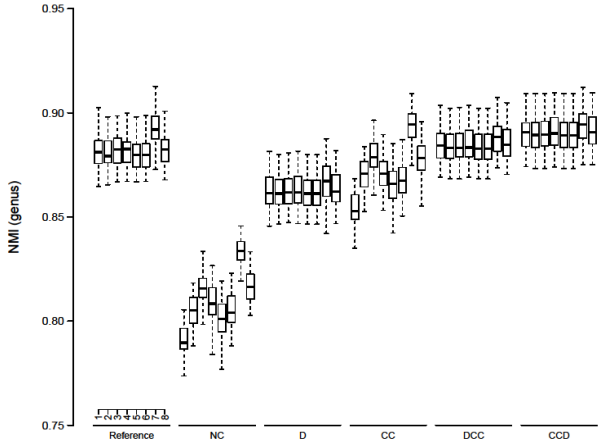


***FIGURE 1**. Genus-level NMI scores by each algorithm after each pre-processing approach are shown for 50 CORE simulations. Reference: reference clustering of non-chimeric and non-erroneous reads, NC: no cleaining, D: denoised, CC: chimera checked, DCC: denoised and chimera checked, CCD: chimera checked and denoised. 1- CD-HIT, 2- DNA-CLUST, 3- ESPRIT-TREE, 4- UCLUST, 5- USEARCH, 6- USEARCH Optimal, 7- CLUSTOM, 8- TBC.*

REFERENCES

1. Griffen, A.L. *et al.* (2011) CORE: a phylogenetically-curated 16S rDNA database of the core oral microbiome. *PLoS ONE*, **6**, e19051.

RAPID IDENTIFICATION OF POTENTIAL ANTIMICROBIAL DRUG TARGETS

Fredrick M. Mobegi[1, 2], Sacha A. F. T. van Hijum[2], Peter Burghout[1], Hester J. Bootsma[1], Stefan P.W. de Vries[1], Christa Gaast-deJongh[1], Elles Simonetti[1], Jeroen Langereis[1], Hendrik G. Stunnenberg[3], Peter W. M. Hermans[1], Marien Jonge[1], and Aldert Zomer[1, 2,]*

Laboratory of Paediatric Infectious Diseases [1], Centre for Molecular and Biomolecular Informatics[2], and Department of Molecular Biology[3], Nijmegen Centre for Molecular Life Sciences, Radboud University Medical Centre, Nijmegen.

[*]a.zomer@cukz.umcn.nl

Respiratory tract infections, mainly caused by *Streptococcus pneumoniae*, *Haemophilus influenzae* and *Moraxella catarrhalis* are among the leading causes of global mortality and morbidity. Increased resistance of these pathogens to existing antibiotics necessitates the search for novel targets to develop potent antimicrobials. We have effectively combine high-density transposon mutagenesis, high-throughput sequencing and integrative genomics, to reliably identify potential drug targets in these human respiratory pathogens. We observe that averagely 20% of all genes in the three species are essential for growth and viability. They include 128 essential and conserved genes, part of 47 metabolic pathways. By excluding genes having homologs in human and commensal human gut microflora, we have defined a set of 249 drug targets, 67 of which are targeted by 75 FDA-approved antimicrobials, and 35 other researched small molecule inhibitors. Four novel targets were experimentally validated. Our approach circumvents the tedious and expensive laboratory screens for drug target selection, therefore, accelerating directed drug discovery.

INTRODUCTION

Elucidation of genes essential for bacterial growth and viability is a prerequisite for identifying potential drug targets. High-throughput transposon insertion sequencing strategies, such as Tn-seq, TraDIS, or variants thereof[1,2], are a method of choice for assaying gene essentiality. Here, we have rationally applied Tn-seq and other *in silico* techniques to reliably identify potential drug targets in *S. pneumoniae, H. influenzae and M. catarrhalis*. Using researched inhibitors, we have experimentally validated four of the finally identified novel drug targets. We believe that the findings in this study will advance the general understanding of respiratory pathogens, and facilitate rapid and cost-effective screening for novel drug targets, potentially leading to the discovery of novel antibiotics to treat RTI.

METHODS

Genome annotations information for *S. pneumoniae* R6, *S. pneumoniae* TIGR4, *H. influenzae* 86 028NP, *H. influenzae* Rd KW20, and *M. catarrhalis* BBH18 were updated using RAST. The proteins with updated annotations were then clustered using OrthoMCL, and their subcellular localizations predicted in various publicly available tools. ESSENTIALS[3] was then used to analyze various in-house and literature transposon mutant libraries, and predict the essentiality metric for each ORF. Comparing the ensuing essential genes with the catalogue of human gut microbial genes[4], as well as with the human genome helped to eliminate genes with conserved homologs, and subsequently prioritize potential drug targets (Figure 1).

RESULTS & DISCUSSION

We have identified 249 potential drug targets, 67 of which are acknowledged targets for 75 FDA-approved antimicrobial drugs and 35 other researched small molecule inhibitors[5]. This study, therefore, establishes a foundation for future research into experimental validation of all possible targets identified hitherto. We anticipate that these studies will eventually provide druggable targets that will be successfully moved to drug development.

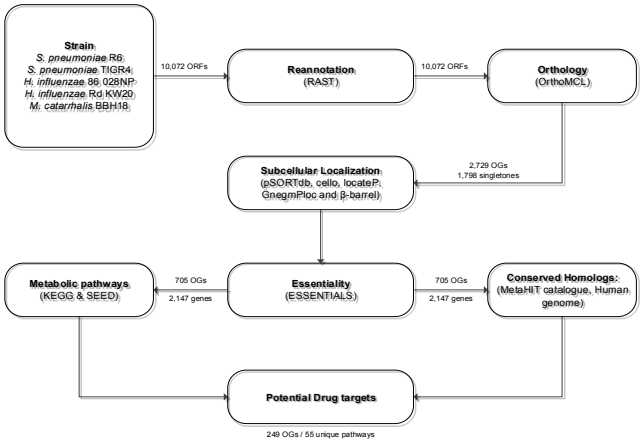


***FIGURE 1**. Schematic overview of the drug targets selection criteria.*

REFERENCES

1. van Opijnen, T., Bodi, K. L. & Camilli. *Nature methods* **6**, 767-772, doi:10.1038/nmeth.1377 (2009).
2. Christen, B. *et al. Mol Syst Biol* **7**, 528, doi:10.1038/msb.2011.58 (2011).
3. Zomer, A., Burghout, P., Bootsma, H. J., Hermans, P. W. & van Hijum, S. A. *PLoS One* **7**, e43012, doi:10.1371/journal.pone.0043012 (2012).
4. Qin, J. *et al. Nature* **464**, 59-65, doi:10.1038/nature08821 (2010).
5. Knox, C. *et al. Nucleic Acids Res* **39**, D1035-1041, doi:10.1093/nar/gkq1126 (2011).

THE COMPLEX INTRON LANDSCAPE AND MASSIVE INTRON INVASION IN A PICOEUKARYOTE PROVIDES INSIGHTS INTO INTRON EVOLUTION

Bram Verhelst[1,2], *Yves Van de Peer*[1,2], *Pierre Rouzé*[1,2,*].
Dept. of Plant Biotechnology and Bioinformatics, Ghent University[1]; *Dept. of Plant Systems Biology, VIB*[2].
[*]*pierre.rouze@psb.vib-ugent.be*

This study focuses on two strains, CCMP1545 and RCC299, and their related individuals from ocean samplings, showing that they not only harbour different classes of introns depending on their location in the genome, as for other Mamiellophyceae, but uniquely carry several classes of repeat introns. These introns, dubbed introner elements (IEs), are found at novel positions in genes and have conserved sequences, contrary to canonical introns. This IE invasion has a huge impact on the genome, doubling the number of introns in the CCMP1545 strain. Along with similar cases recently observed in other organisms, our observations in Micromonas strains shed a new light on the evolution of introns, suggesting that intron gain is more widespread than previously thought.

INTRODUCTION

The Mamiellophyceae include eukaryotic picoalgae at the basis of the green linear that play a major trophic role in the marine environment. Amongst these are two Micromonas strains (CCMP1545 and RCC299) with unique intron landscapes (Figure 1). In common with other Mamiellophyceae, the heterogeneic genome results in the presents of two distinct intron classes: canonical and BOC1. Additionally to these classes (and unlike the other sequenced Mamiellophycae genomes), repeat introns were discovered and dubbed Introner Elements (IEs).

RESULTS & DISCUSSION

There are four IE classes (three in CCMP1545 and one in RCC299), and they differ in terms of host, abundance, sequence and length. These introner elements are always found at novel positions within genes (in contrast to canonical introns), are located in the sense strand of genes, and display proper splicing signals.

IEs are not evenly distributed in the genome and are virtually absent from low-GC% areas (Figure 1). There are no sequence motifs or nearby signals that explain the presence of IEs within genes.

Using metagenomic sequences and a close relative of RCC299 (CCMP1764), we discovered presence/absence polymorphisms (PAPs), wherein the Micromonas genome would carry an IE at a given locus and the metagenomic sequence did not, or vice versa. In total, we ended up with 511 novel IE positions. This data suggests that IEs are mobile elements that replicate themselves and transpose into new locations. As they are only found in genes in the sense orientation, their mobility is likely linked to the transcription/splicing process. The mechanisms most likely to explain this scenario is known as 'intron transposition', whereby an IE will invade a gene transcript through reverse splicing and subsequent homologous recombination.

The amplitude of the intron gain, coupled to the relative low number of resident introns makes this a case a true advocate for the intron-late scenario. Along with similar cases recently observed in other organisms, our observations in Micromonas strains shed a new light on the evolution of introns, suggesting that intron gain is more widespread than previously thought.

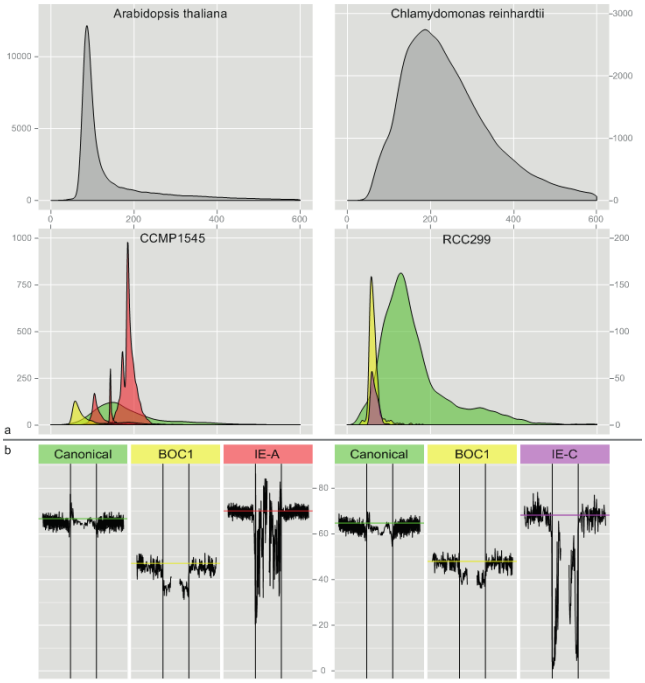


***FIGURE 1**. The Intron landscape of Micromonas. A) Intron Size distribution. B) Average GC content of Micromonas introns and their bordering exon regions (vertical black lines denote the exon/intron boundaries). Intron classes: canonical (green), BOC1 (yellow), IE-A (red) and IE-C (purple).*

REFERENCES

1. Worden *et al. Science* **324**, 268-272 (2009).
2. Moreau *et al. Genome Biology* **13**, R74 (2012).

TOXIN-ANTITOXIN MODULE DYNAMICS CAN CAUSE PERSISTER CELL FORMATION IN *E. COLI*

Alexandra Vandervelde[2,3,], Lendert Gelens[1], Lydia Hill[1,2], Jan Danckaert[1], Remy Loris[2,3]*

*Applied Physics Research Group (APHY), Vrije Universiteit Brussel, Belgium[1]; Molecular Recognition Unit, Department of Structural Biology, VIB, Brussels, Belgium[2], Structural Biology Brussels, Department of Biotechnology, Vrije Universiteit Brussel, Belgium[3] *alexandra.vandervelde@vub.ac.be*

In several chronic infectious diseases, bacterial persister cells play an important role. Persisters are tolerant to multiple antibiotics because they are in a dormant state. This dormancy can be caused by toxin-antitoxin (TA) modules. TA modules are small genetic elements which are widespread on bacterial genomes. They encode two proteins: a toxin that slows down or halts the bacterial metabolism, and an antitoxin that is able to neutralize this toxin. The transcription of TA modules is regulated by toxin-antitoxin complexes, binding on the operator depending on the relative amounts of toxin and antitoxin present. In order to unravel this sophisticated regulation and the role of TA modules in persister cell formation, we built stochastic models describing these biological networks. As it turns out, an increase in the number of binding sites on the operator allows a more economical maintenance of the TA module by decreasing the protein production. Furthermore, we were able to simulate the formation of persisters through rare increases in the free toxin level. Therefore, we believe that our models can contribute to our understanding of TA regulation and its relation to the generation of persister cells.

INTRODUCTION

Toxin-antitoxin (TA) modules are small genetic elements, widespread on bacterial genomes and plasmids, which code for an intracellular toxin and its corresponding antitoxin. Although their biological function is still debated, evidence accumulates that TA modules are implicated in the bacterial stress response[1] and the formation of persisters, cells that are tolerant to environmental stresses such as antibiotics because they are in a dormant, non-dividing state[2]. The negative transcriptional regulation of TA modules relies on conditional cooperativity: the toxin can function as a corepressor or a derepressor for the DNA-binding antitoxin, depending on the toxin:antitoxin ratio. We captured this autoregulation in two models: one simplified model, in which the binding sites on the operator behave independently (Figure 1B), and a model including more complex binding processes on the DNA (Figure 1C).

METHODS

The mathematical models are based on published structural, molecular and biophysical data for toxin-antitoxin modules, summarized in Figure 1. The model parameters are based on the *ccd* toxin-antitoxin system. Stochastic simulations were performed using a Gillespie algorithm[3].

RESULTS & DISCUSSION

Simulations based upon the model with independent binding sites display the expected characteristics for TA modules, like a low free toxin level and high antitoxin levels in non-starvation conditions.

Using this model, we further found that an increase in the number of binding sites on the operator of a TA module leads to decreased protein levels and a decreased variability for the free antitoxin and complexes AT and TAT, while the free toxin level stays low and relatively constant. This decrease in the protein concentrations allows a more economical maintenance of the toxin-antitoxin system, which is an advantage for the cell.

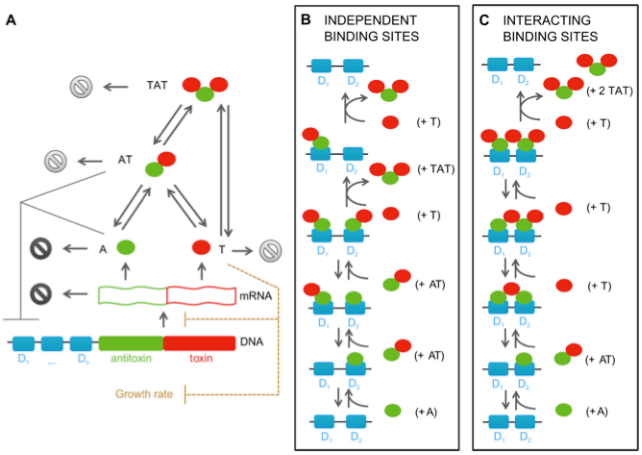


***FIGURE 1**. Toxin-antitoxin models for one or more binding sites on the operator, based on the repression.*

Finally, using the model with interacting binding sites on the operator, rare, extreme stochastic spikes in the free toxin levels were found. These spikes provide a route to persister generation, as the presence of free toxin decreases the growth rate of a cell. By including this toxic feedback effect in our model, we found that the duration of the persister state is closely related to the amplitude of the toxin spike.

REFERENCES

1. Gerdes K *et al.* Prokaryotic toxin-antitoxin stress response loci. *Nat Rev Microbiol* **3**: 371–382 (2005).
2. Lewis K. Persister cells, dormancy and infectious disease. Nat Rev Microbiol **5**: 48–56 (2007).
3. Gillespie D. Exact stochastic simulation of coupled chemical-reactions. *Journal of Physical Chemistry* **81**: 2340–2361 (1977).
4. Gelens, L., Hill, L., Vandervelde, A., Danckaert, J. & Loris, R. A general model for toxin-antitoxin module dynamics can explain persister cell formation in *E. coli*. PLoS Comput. Biol. **9** (2013).

WHOLE-GENOME SEQUENCE-BASED IDENTIFICATION OF EPIDEMIC PLASMIDS SPREADING EXTENDED-SPECTRUM BETA-LACTAMASE GENES AMONG *ESCHERICHIA COLI* FROM DIFFERENT HOSTS

Mark de Been[1,*], *Val Fernandez*[2], *María de Toro*[2], *Jelle Scharringa*[1], *Yu Du*[3], *Juan Hu*[3], *Zhangyi Liu*[3], *Ying Lei*[3], *Zhong Chen*[3], *Ad Fluit*[1], *Marc Bonten*[1], *Rob Willems*[1], *Fernando de la Cruz*[2], *Willem van Schaik*[1].

Dept. of Medical Microbiology, University Medical Centre Utrecht (UMCU), the Netherlands[1]*; Dept. of Molecular Biology, School of Medicine, Universidad de Cantabria & Instituto de Biomedicina y Biotecnología de Cantabria, Santander, Spain*[2]*; BGI-Shenzhen, Shenzhen, China*[3]*. *M.deBeen-2@umcutrecht.nl*

Incidences of infections caused by extended-spectrum β-lactamase (ESBL)-producing Enterobacteriaceae are rapidly increasing worldwide and results from recent studies have suggested that these bacteria, which are also highly prevalent among food-producing animals, can spread clonally through the food chain to humans. Here, we used whole-genome sequencing (WGS) to study the relatedness of ESBL-producing *Escherichia coli* from humans, chicken retail meat, chickens and pigs. Our data suggest that ESBL genes are mainly disseminated through the food chain via epidemic plasmids.

INTRODUCTION

Recent studies performed in the Netherlands demonstrated that humans, retail chicken meat and chickens share the same ESBL genes that are encoded on similar plasmid backbones in *E. coli* strains with similar multi-locus sequence types (MLST), suggesting clonal transfer of ESBL-producing *E. coli* from poultry to humans[1-3]. However, these interpretations were based on the results of classical typing methods that target only a limited number of house-keeping and virulence genes, and which may not provide sufficient resolution to accurately monitor the epidemiology of ESBL-producing bacteria. In this study, we have, therefore, whole-genome sequenced 32, mostly ESBL-producing, *E. coli* strains from humans, retail chicken meat, chickens and pigs from two studies. One included pairs of human and poultry-associated strains that had previously been considered indistinguishable based on strain, plasmid and ESBL gene typing. The other included isolates from pigs and farmers from the same farm.

METHODS

The 32 *E. coli* strains sequenced in this study had been isolated from humans (n=17), chicken (n=4), chicken meat (n=7) and pigs (n=4) in the Netherlands between 2006-2011. The strains included three different STs (ST10, n=7; ST58, n=4; ST117, n=3) and four different ESBL genes ($bla_{CTX\text{-}M\text{-}1}$, n=20; $bla_{CTX\text{-}M\text{-}15}$, n=1; $bla_{TEM\text{-}52}$, n=5; $bla_{SHV\text{-}12}$, n=2), which had previously been found on an IncI1/ST7 (n=8), IncI1/ST10 (n=1) or IncI1/ST36 (n=2) plasmid. This set also contained four non-ESBL-producing strains, which were included as controls.

Strains were sequenced using Illumina HiSeq 2000 sequencing technology generating 90 bp paired-end reads from a library with an average insert size of 500 bp and a total amount of quality-filtered raw sequence of over 600 Mb per strain. Genomes were assembled *de novo* using SOAPdenovo[4].

RESULTS & DISCUSSION

Phylogenomic and core genome analyses revealed considerable heterogeneity between human and poultry-associated isolates. The most closely related pairs of strains from both sources, which had previously been found to be identical on the basis of strain, plasmid and ESBL gene typing, carried over 1 kb SNPs per Mb core genome. In comparison, epidemiologically linked isolates from humans and pigs differed by only 1.8 SNPs per Mb core genome.

To investigate the possibility of horizontal spread of ESBL genes via ESBL-carrying plasmids through the *E. coli* population, we used a novel approach to reconstruct plasmids from WGS data. Application of this approach resulted in the reconstruction of 148 plasmids (average of 4.6 ± 2.1 plasmids per strain), with plasmid sizes ranging from 1.1 kb to 290.4 kb (median size of 5.8 kb). There was excellent agreement between previously obtained plasmid typing data and WGS-based plasmid reconstructions. Among the reconstructed plasmids, we found two sub-families of virtually identical ESBL-carrying IncI1 plasmids, one corresponding to IncI1/ST3 (n=6; 0-4 SNPs / 40 kb plasmid core) and one corresponding to IncI1/ST7 (n=12; 0 SNPs / 50 kb core). We also identified a sub-family of nearly identical AmpC-type β-lactamase-carrying IncK plasmids (n=9; 0-27 SNPs / 37 kb core). All three plasmid types occurred in genetically unrelated human, poultry and/or pig isolates, suggesting that they efficiently spread through the *E. coli* population and play an important role in the dissemination of ESBL- and AmpC-type β-lactamases.

In conclusion, our data failed to confirm the evidence for clonal transmission of ESBL-producing *E. coli* from poultry to human, as has been suggested based on classical typing methods. Instead, our data suggest that ESBL genes are mainly disseminated via epidemic plasmids.

REFERENCES

1. Kluytmans *et al. Clin Infect Dis* **56**, 478-487 (2013).
2. Leverstein-van Hall *et al. Clin Microbiol Infect* **17**, 873-880 (2011).
3. Overdevest *et al. Emerg Infect Dis* **17**, 1216-1222 (2011).
4. Li *et al. Hum Genomics* **4**, 271-277 (2010).

EVOLUTION FOLLOWING WHOLE GENOME DUPLICATION IN THE YEAST GENE REGULATORY NETWORK

Ehsan Sabaghian[1,*], *Pieter Meysman*[2], *Kris Laukens*[2], *Bart Goethals*[3], *Riet De Smet*[1], *Yvan Saeys*[4], *Yves Van de Peer*[1].

Department of Plant system biology, division of Bioinformatics and Systems Biology, Ghent University[1]*; Biomedical informatics research center Antwerp (biomina)*[2]*; Department of Mathematics and Computer Sciences, Antwerp University*[3]*; Department of Molecular Biomedical Research, Ghent University*[4]*. *ehsan.sabaghian@psb.vib-ugent.be*

Genome duplication plays an important role in the evolution of organisms. Because genome-wide duplication events duplicate whole molecular networks it is of interest to investigate how these networks evolve subsequent to such events. Access to huge amounts of omics data make such analyses possible.

INTRODUCTION

Gene duplication can be defined as any duplication of a region of DNA that contains a gene. Duplicated genes are considered of major importance for evolutionary novelty since they provide the necessary increase in raw genetic material on which evolution can work. In particular whole genome duplication (WGD) events provide a major source of duplicate genes. Duplicated genes can contribute to functional innovation by evolution of their coding sequences, yet duplication might also trigger rewiring of the regulatory network. Here we will focus on the latter and investigate how an ancient WGD event within yeast has influenced the evolution of the gene regulatory network.

METHODS

We considered a curated yeast gene regulatory network[1]. This network is directed and consists of 6405 genes and 48082 interactions. We will focus on 834 genes in this network which have been retained after a whole genome duplication that occurred around 100 million years ago[2,3].

First, we explored the network topological properties of the genes duplicated by WGD. We extracted different topological properties from the gene regulatory network and used Random Forests to test whether these are different for duplicated genes as compared to non-duplicated genes.

In a second approach we verified how WGD has affected the evolution of transcription factors. In specific we assessed the divergence in their upstream regulatory genes as well as their downstream target genes. To this end we selected 22 duplicated transcription factor pairs and compared the overlap in their regulatory genes and target genes with that of a set of 16631 randomly selected gene pairs. The hypergeometric test was used to assess the significance in overlap of the gene sets.

P-value of hypergeometric test for testing a significant overlap between two pairs considered as a comparison criterion.

RESULTS & DISCUSSION

Topological properties

Using a large set of topological properties as variables the error rate in the detection of duplicated genes by Random Forests is very high (96%). This was to be expected since the set of duplicated genes considered is highly heterogeneous in their topological properties. However, features related to network centrality are more informative in distinguishing duplicated.

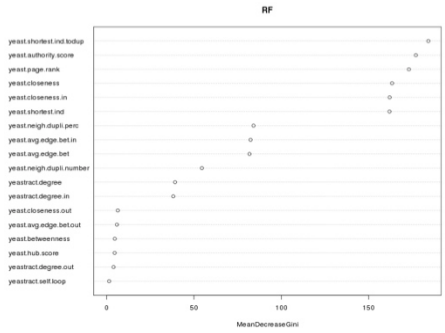


***FIGURE 1**. Variable importance in Random Forests.*

Analysis of transcription factor pairs

The comparison of the overlap in target genes and regulatory genes for duplicated transcription factor pairs and randomly selected gene pairs revealed that the duplicated transcription factors have a significantly higher overlap in both target genes and regulatory genes than the randomly selected gene pairs. This suggests that the duplicated transcription factors have not entirely diverged yet in their regulation following WGD.

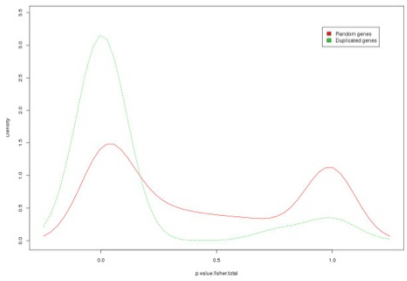


***FIGURE 2**. P-value distribution for the overlap in target genes.*

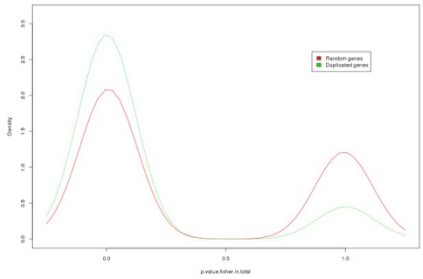


***FIGURE 3**. P-value distribution for the overlap in regulatory genes.*

REFERENCES

1. Abdulrehman D *et al*. *Nucl. Acids Res* **39**, D136-D140 (2011).
2. Manolis Kellis *et al*. *Nature* **428**, 617-624 (2004).
3. Kenneth H *et al*. *Nature* **387**, 708-713 (1997).

COMPARATIVE TRANSCRIPTOMICS OF HELPER T CELLS

Henk-Jan van den Ham[1,*], *Fatiha Zaaraoui*[1], *Wilfred F van IJcken*[2], *Albert D Osterhaus*[1], *Rob J de Boer*[3] & *Arno C Andeweg*[1].

Depts. of Viroscience[1] *and Biomics*[2], *Erasmus Medical Center, Rotterdam ; Theoretical Biology & Bioinformatics, Universiteit Utrecht*[3]. [*] *h.j.vandenham@erasmusmc.nl*

Helper T (Th) cells are important producers of cytokines that regulate the immune system. Th cells have the capacity to adopt different phenotypes, i.e., produce dichotomous sets of effector cytokines. We examine dense time-course transcriptome snapshots of mouse and human Th cell differentiation to identify similarities and differences between Th cells in these species. We find that gene expression is generally conserved, but that there is a difference in kinetics between mouse and human.

INTRODUCTION

Helper T cells are important orchestrators of the immune response that regulate through the production of cytokines. Over the past decade, it has been shown that, both in mouse and man, CD4+ helper T (Th) lymphocytes have the capacity to differentiate into a range of different phenotypes by expressing distinct sets of cytokines[1]. These cytokines determine effector mechanism employed by the immune response, such as cellular (cytotoxic T cell) or humoural (antibody) immunity. Upon activation, naïve Th cells are skewed toward a particular phenotype by their environment and form Th clones by several rounds of cell division. Many established markers of Th cell phenotypes are conserved between mice and humans, but mouse-man differences have also been documented. Here, we perform a comprehensive comparison of mouse and human Th cell differentiation through detailed analysis of dense time-course transcriptome snapshots.

METHODS

In this study, we apply principle component and correlation analysis, limma[2], polar score analysis[3], and weighted gene correlation network analysis (WGCNA)[4] to examine Th transcriptome conservation between mice and humans in dense time-course experiments. We extend limma and the polar score procedure to enable analysis of dense time-course data. Mouse and human expression data were reprocessed and reannotated using ensembl gene-based probeset definitions[5] to incorporate the most recent genome information and enabling one-to-one comparisons of the expression level of human and mouse orthologous genes.

RESULTS & DISCUSSION

We have previously shown in mice that, at the transcriptome level, Th cell activation leads to much larger changes in gene expression than does Th phenotype skewing[6]. Here, we reiterate this result for both mouse and human data. Global comparisons of dense time-course datasets reveal that Th cell transcriptomes of mouse and human indeed show a large degree of concordance, in particular for genes that change over the course of Th differentiation. Furthermore, we see a difference in kinetics between mouse and human Th cells, which contributes to the differences in gene expression between mouse and human regulation in addition to species-specific differences in Th regulation between mouse and human. Co-expression network analysis indicates that modules related to cell activation and to Th2 cell differentiation are conserved.

In summary, we find that helper T cell gene expression in orthologous gene pairs is generally conserved between mouse and human, regardless of the analysis technique used. Furthermore, we find that mouse cells differentiate more slowly than human Th cells.

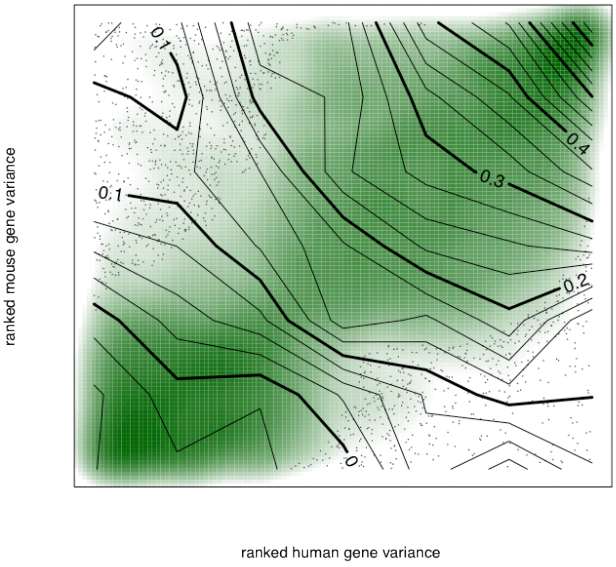


***FIGURE 1**. Correlations between orthologous genes change with human and mouse gene variance.*

REFERENCES

1. Van den Ham, H.-J. *et al. Par Immunol* 1–13 (2013). doi:10.1111/pim.12053
2. Smyth, G. K. *Statistical applications in genetics and molecular biology* **3,** Article3 (2004).
3. Van den Ham, H. J. *et al. J Immunol Methods* **361,** 98–109 (2010).
4. Langfelder, P. & Horvath, S. *BMC Bioinformatics* **9,** 559 (2008).
5. Dai, M. *et al. Nucleic Acids Res* **33,** e175 (2005).
6. Van den Ham, H.-J. *et al. Eur J Immunol* **43,** 1074–84 (2013).

EFFECT OF UNBS1450 ON HISTIOCYTIC LYMPHOMA CELL LINE U937: A TRANSCRIPTOMIC ANALYSIS

Gaigneaux Anthoula[*], *Chateauvieux Sebastien, Cerella Claudia, Dicato Mario, Diederich Marc.*
Laboratoire de Biologie Moléculaire et Cellulaire du Cancer (LBMCC). Hôpital Kirchberg, 9, rue Edward Steichen, L-2540 Luxembourg. [*]*anthoula.gaigneaux@lbmcc.lu*

INTRODUCTION

UNBS1450 is a hemisynthetic cardenolide, chemically modified form of 2-oxovoruscharin, a derivative of voruscharin, extracted from Calotropis procera, a plant widely used in traditional medicine. Our previous results demonstrated that UNBS1450 has a dose-dependent anti-tumor effect, inhibiting proliferation in various cancer cell lines and inducing apoptotic cell death[1-3].

The aim of this work is to evaluate the effect of UNBS1450 on the transcriptome of the histiocytic lymphoma cell line U937. By this approach we intend to highlight novel pathways and species targeted by this treatment, especially focusing on early time points, i.e. prior to induction of apoptosis.

METHODS

We analyzed transcriptional effect of UNBS1450 after 3, 6 and 9 hours of treatment at 20 nM, an apoptogenic concentration as previously showed. We assessed transcriptional effect using dual-color microarrays. Gene foldchanges and lists of differently expressed genes were obtained using Limma (from R/Bioconductor environment). Significant GO categories associated with 9h gene list were retrieved using the Bioconductor package Gostats[4]. Pathway enrichment results for 3h and 6h were obtained using GSEA software[5]. For both kind of analyses, we kept only groups whose size was comprised between 15 and 500 genes, and $p<0.01$ was used as statistical threshold. Bioinformatic results were visualized in Cytoscape using Enrichment Map plug-in[6].

RESULTS & DISCUSSION

Our results show that UNBS1450 had only a reduced transcriptional effect at early time points, with 50 and 41 significant genes after 3h and 6h of treatment, respectively. Global foldchange between conditions was enlarging with time and the effect at 9 h was stronger, with 744 significantly affected genes. Several bioinformatics tools were used to summarize results in biological categories. Using a combination of gene set analysis and network display, we have shown important pathways and biological processes affected by UNBS1450 treatment at several timepoints prior to apoptosis. Especially, effects on mitosis and histones were shown at earlier time points. At 6h, our results have shown effects on ontologies related to histones, ribosomes, translation and tubulins. After 9 hours of treatment, several GO categories related to inflammation and cytokines, as well as differentiation, were regulated. Using global enrichment analyses, we have shown important pathways and biological processes affected by UNBS1450 treatment. These results will be further validated by wet lab analyses.

REFERENCES

1. Juncker, T., et al., UNBS1450 from Calotropis procera as a regulator of signaling pathways involved in proliferation and cell death. Biochemical pharmacology, 2009. 78(1): p. 1-10.
2. Juncker, T., et al., UNBS1450, a steroid cardiac glycoside inducing apoptotic cell death in human leukemia cells. Biochemical pharmacology, 2011. 81(1): p. 13-23.
3. Cerella, C., M. Dicato, and M. Diederich, Assembling the puzzle of anti-cancer mechanisms triggered by cardiac glycosides. Mitochondrion, 2013. 13(3): p. 225-34.
4. Falcon, S. and R. Gentleman, Using GOstats to test gene lists for GO term association. Bioinformatics, 2007. 23(2): p. 257-8.
5. Subramanian, A., et al., GSEA-P: a desktop application for Gene Set Enrichment Analysis. Bioinformatics, 2007. 23(23): p. 3251-3.
6. Merico, D., et al., Enrichment map: a network-based method for gene-set enrichment visualization and interpretation. PLoS One, 2010. 5(11): p. e13984.

RNA-SEQUENCING IDENTIFIES NOVA1 AS A MAJOR SPLICING REGULATOR IN PANCREATIC BETA CELLS

Olatz Villate, Jean-Valery Turatsinze*, Tatiane C. Nogueira, Fabio A. Grieco & Decio L. Eizirik.*

Laboratory of Experimental Medicine, Medical Faculty, Université Libre de Bruxelles, Brussels (ULB), Belgium.

**olatzv@gmail.com, jturatsi@ulb.ac.be*

The control of RNA alternative splicing is critical for generating biological diversity. The advent of whole transcriptome sequencing (RNA-seq) technologies enabled the study of alternative splicing at the genome wide level. Here we present the results of RNA-seq of rat pancreatic beta cells following the knock-down of the Nova1 splicing factor. Our results suggest that Nova1, previously considered a brain specific protein, is a major splicing regulator in pancreatic beta cells and identify several splicing events in the rat transcriptome.

INTRODUCTION

Alternative splicing (AS) is a basic mechanism for the regulation of gene expression, affecting more than 90% of human genes. We have recently shown by exon array and RNA-sequencing that human pancreatic beta cells exposed to the pro-inflammatory cytokines interleukin-1 beta (IL-1β) + interferon-gamma (IFN-γ) show changes in AS of >500 genes. Our recently published data[1] indicate the presence of the Nova1 splicing factor in beta cells, but there is no information about its role on these cells.

METHODS

Nova1 expression was inhibited by 60% with the use of two independent siRNAs (knockdown, KD). Three independent preparations of FACS-purified primary rat beta cells (90-95% pure) were RNA-sequenced in control conditions or after 48-hour KD of Nova1. Samples were sequenced on an Illumina HisSeq 2000 sequencer. After quality assessment, the reads were mapped to the rat genome using Tophat[2] mapper. The transcript assembly, abundance estimation, differential expression and splicing were performed with the FluxCapacitor[3] (http://flux.sammeth.net) software suite using the Ensembl rat annotations dataset. Expression or splicing was considered changed if it fulfilled two criteria: 1. A Benjamini-Hochberg-corrected Fisher-test p-value < 0.05; 2. A modification in the same direction in all samples. The percentage of splicing index score was calculated to evaluate the level of exon inclusion/exclusion. The transcript *de novo* assembly was performed using Cufflinks[3] tools to identify novel isoforms.

RESULTS & DISCUSSION

RNA-sequencing and the subsequent bioinformatics analysis identified 24,162 transcripts as expressed in primary beta cells (RPKM ≥ 0.5), corresponding to 12,723 genes. Nova1 KD modified expression of 26% of these genes, altering the splicing of 3452 of them.

Using the number of identified splice junctions we observed 10-fold more splicing events (~24000 events) as compared to the observed splicing events observed with the current rat annotations (2400 events).

Pathway analysis suggested that the modified genes are involved in exocytosis, apoptosis, lipid metabolism, splicing and transcription. The effects of Nova1 KD on the splicing of key downstream genes such as Gabrg2 and Neurexin1 were confirmed by RT-PCR. In line with the observed pattern of gene expression, Nova1 silencing inhibited insulin secretion but did not affect glucose oxidation. Importantly, Nova1 silencing induced apoptosis in INS-1E cells and FACS purified rat beta cells, basally and after cytokine treatment.

The present data provide the first indication that Nova1 has a major role in beta cell function, controlling the splicing and expression of key genes involved in exocytosis and apoptosis. These findings provide relevant information to better understand the role of AS in pancreatic beta cells

Grant acknowledgement: *This work was supported by grants from the JDRF, the Fonds National de recherché Scientifique (FNRS) and the European Union (project BetaBat).*

.

REFERENCES

1. Eizirik DL *et al*. *Plos Genet Opin* **8**, e100255 (2012).
2. Trapnell C *et al. Bioinformatics* **25** 1105-1111 (2009).
3. Montgomery SB *et al. Nature* **464**: 773-777 (2010).
4. Trapnell C *et al. Nat. Biotechnol.* **28** 511-515 (2010).

CELLMISSY: A TOOL FOR MANAGEMENT, STORAGE, DISSEMINATION AND ANALYSIS OF CELL MIGRATION DATA

Paola Masuzzo[1,2], *Niels Hulstaert*[1,2], *Lynn Huyck*[2], *Christophe Ampe*[2], *Marleen Van Troys*[2] & *Lennart Martens*[1,2,*].

Department of Medical Protein Research, VIB[1], *and Faculty of Medicine and Health Sciences*[2], *Department of Biochemistry, Ghent University.* [*]*lennart.martens@ugent.be*

Automated image acquisition and quantification strategies, together with the emergence of higher throughput cell migration assays (e.g. in relation to *in vitro* drug screening), is inciting cell migration to evolve to a high-throughput research field. However, data from different sources still need to be manually transferred across multiple downstream data analysis tools. As a consequence, the current challenges in the field lie in data management, data storage and streamlining data (meta)analysis. We here present CellMissy, a cross-platform tool which simplifies and automates data management, storage, dissemination and analysis of cell migration data, from the initial experimental set-up to final data exploration and analysis.

INTRODUCTION

Cell migration is essential for morphogenesis, immune response, wound healing and cancer metastasis[1]. Investigation of the mechanisms and modes of this process is therefore important for fundamental scientific insight and translational research. As a result, the field has developed methods for the high-throughput acquisition of cell migration data and a major challenge currently lies in the development of suitable software for handling these data[2]. Although software tools are available to help with several of the steps in a typical experimental workflow, several key functions remain unmet. We here present CellMissy, a cross-platform system for the annotation, storage, dissemination and analysis of quantitative cell migration data[3].

METHODS

CellMissy is a Java client-server application with a graphical user interface on the client and a relational database in the back-end. Developed to follow the steps typically encountered in a cell migration experiment, it is composed of three modules: the experiment manager, the data loader and the data analyzer. At several points, PDF and XML files can be generated to assist reporting and data dissemination. CellMissy architecture is fully pluggable for analysis methods, which means that new analysis algorithms can be added by any interested developer.

RESULTS & DISCUSSION

The first CellMissy module is used to set up a cell migration experiment: the user defines technical replicates and biological conditions through a multiwell plate and by choosing different informative metadata variables (e.g. cell line, treatment). The experimental set-up can be exported as PDF file and as XML template, which allows sharing a set-up between laboratories, thus promoting data exchange and reproducible research. Data import and storage are handled by the data loader module, which either reads a generic text-based data input format or can be fully automated if tailored to the chosen combination of set-up/image processing tool[4]. CellMissy again allows exporting the experiment to an XML file at this level, i.e. with all acquired raw data. Finally, the data analyzer module guides the user through data inspection, analysis and interpretation. Data quality control includes among others an assessment of technical precision between replicates (Figure 1). Statistical analyses comparing biological conditions can be executed, enabling quantitative statements on differences in cell migration features across biological conditions. Finally, CellMissy produces a detailed and customizable PDF report that summarizes experimental details, acquired data, corresponding results and statistical evaluation. In conclusion, CellMissy is a novel, cross-platform, generic data management, dissemination and analysis system for cell migration experiments. It is freely available at http://cellmissy.googlecode.com, easily extensible and automates data handling, processing, quality monitoring, sharing and interpretation.

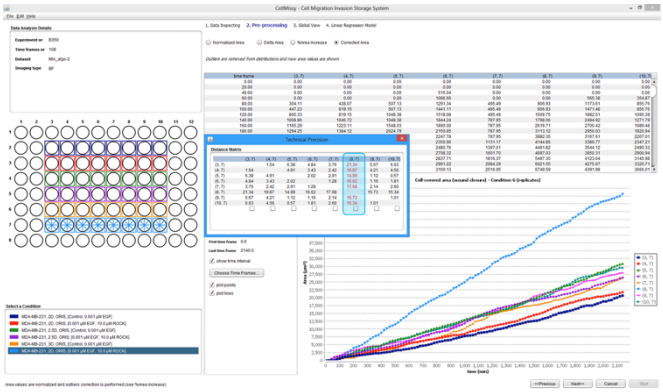


***FIGURE 1**. Part of the data analyzer module: assessment of the technical precision within a biological condition.*

REFERENCES

1. P. Friedl and D. Gilmour. Collective cell migration in morphogenesis, regeneration and cancer. *Nat Rev Mol Cell Biol*, (2009) **10**(7): 445-57.
2. Le Dévédec et al. Systems microscopy approaches to understand cancer cell migration and metastasis. *Cell Mol Life Sci*. 2010, **67**(19):3219-40.
3. Masuzzo *et al*. CellMissy: a tool for management, storage and analysis of cell migration data produced in wound healing-like assays. *Bioinformatics* 2013, doi:10.1093/bioinformatics/btt437
4. L. Huyck *et al*. (*in preparation*) Shifting Quantitative Analysis of Migration Dynamics in 3D-matrices to Higher Throughput.

GWAS-M: GENOME-WIDE ASSOCIATION STUDIES FOR MICROBES

Jumamurat R. Bayjanov[1,2], Lennart Backus[1,2], Bas E. Dutilh[1] & Sacha A.F.T. van Hijum[1,3,4].*
Centre for Molecular and Biomolecular Informatics, Radboud University Medical Centre[1]; TI Food and Nutrition, The Netherlands[2]; NIZO food research, Kluyver Centre for Genomics of Industrial Fermentation, The Netherlands[3]; Netherlands Bioinformatics Centre[4]. [*]J.Bayjanov@umcn.nl

Complete or draft genome sequences for microbial organisms are increasing rapidly. They are valuable resources for genotype–phenotype association analysis and gene function prediction, provided that phenotypes are consistently annotated for all the sequenced strains. Here we address the requirements for successful genotype-phenotype association analysis as well as outline a basic protocol for microbial functional genomics using association analysis. The methodologies for genotype-phenotype association analysis described herein can also be applied to other data types.

INTRODUCTION

The recent drop in the cost of genome sequencing has enabled an increase in the scale of comparative genomic analyses. Several thousand bacterial genomes have been sequenced thus far, opening up the potential for microbial genome-wide association studies (GWAS). However, the major bottleneck for applying GWAS to microbial datasets is the scarcity of experimentally consistent phenotypes or other annotations, such as environmental parameters (metadata)[1]. Thus, here we focus only on genotype–phenotype association analysis using machine learning, but the methods outlined herein are equally suitable for finding associations between other data types (e.g.: transcriptome and phenotype data or microbial population structure and environmental parameters).

METHODS

Genotype-phenotype associations can be found using different methods, for example, by comparing sample means or using machine learning algorithms. Regardless of the chosen method, GWAS analysis generally consists of the following three steps (see Figure 1):

- Data preparation.
- Association analysis.
- Interpretation of identified associations.

In data preparation step, genes from different organisms are grouped into groups of orthologous genes. After data preprocessing, the Random Forest machine learning algorithm is used to find gene-phenotype associations due to its versatile property of handling different data types and data sizes[2]. Generally many genotype-phenotype relations are found using this method[3,4], which may include both false and true positive links. Thus, identified associations are visualized for faster screening and better interpretation[3].

RESULTS & DISCUSSION

We used the GWAS-M approach in several studies where genotype information was obtained using DNA microrarrays[3,4] as well as using complete genomic sequences (ongoing work). However, one of the disadvantages of the static measures of genotype (e.g. gene presence/presence) is that they do not take into account other levels of cellular regulation, such as gene expression and protein abundance. Further improvements to such studies could include visualization of identified links within the context of a biological system, for example, in a metabolic network.

The GWAS-M might also be used to link metagenomic entities (e.g.: taxonomy information) observed across metagenomic samples to clinical or environmental metadata. However, extracting taxonomic/function information from metagenomic data is not straightforward.

Future microbial association studies will require the integration of consistent genome annotations with consistent phenotypic datasets. Though novel methods are necessary for handling large (meta)genomic data, the challenge does not lie in the generation of sequence data or in the development of novel statistical techniques, but rather in the generation, annotation and storage of phenotypic data about the genomes and metagenomes that are being studied.

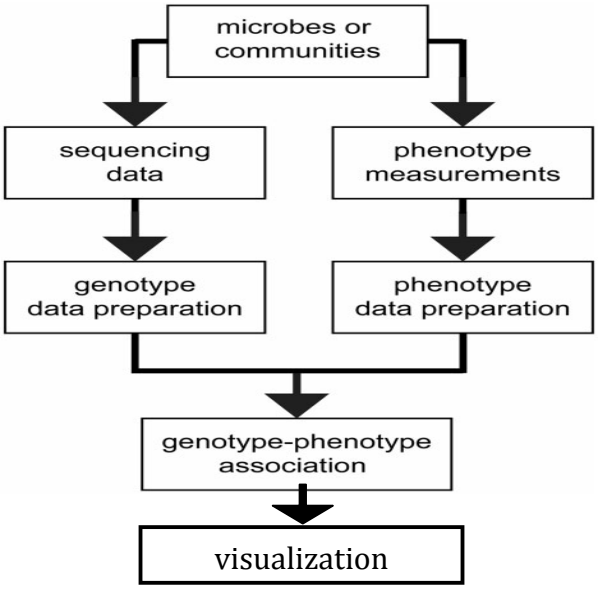


***FIGURE 1**. Flow diagram for genotype-phenotype association analysis.*

REFERENCES

1. Dutilh B *et al. Brief. Funct. Genomics* **4**, 366-380 (2013).
2. Breiman L. *Machine Learning* **45**, 5–32 (2001).
3. Bayjanov JR *et al. BMC Genomics* 13:170 (2012).
4. Bayjanov JR *et. al. BMC Microbiol.* 13:68 (2013).

afkSNP: Assembly-Free K-mer based SNP Comparison of Bacterial WGS Samples

Jeroen Van Goey [1,*], *Hannes Pouseele*[1], *Philip Supply*[2, 3] & *Stefan Niemann*[4]
Applied Maths NV, Sint-Martens-Latem, Belgium[1], *Genoscreen, Lille, France*[2], *INSERM, U1019, CNRS UMR 8204, Institut Pasteur de Lille, Univ Lille Nord de France, Lille, France*[3], *National Reference Center for Mycobacteria, Forschungszentrum Borstel, Borstel, Germany*[4] *info@applied-maths.com

Molecular surveillance of bacterial pathogens as well as bacterial population studies are more and more relying on whole-genome sequencing for high-resolution performance. Fast and reliable strain comparison is of paramount importance in assisting timely epidemiological action, allowing to quickly discriminate outbreak related bacteria from more distantly related strains, independent of their evolutionary speed. The afkSNP method presented here is fast, does not rely on the use of reference sequences and is very well adopted to reflect the individual speed of bacterial evolution.

Introduction

The democratization of next-generation sequencing creates new possibilities for molecular surveillance of bacterial pathogens. However, easy to use bioinformatics tools to analyze these samples are currently lacking, both for technical and biological reasons. On a technical level, routine labs seem to have difficult access to the bioinformatics resources required to analyze (confidential) whole-genome sequencing (WGS) data. On the biological level, for numerous micro-organisms there are no closely related reference sequences available, compromising the resolution of the traditional reference sequence-based approaches. Here we present a reference-free, assembly-free method to compare WGS samples based on single-nucleotide polymorphisms (SNPs). The method is available in the BioNumerics® software, version 7.5 (Applied Maths NV).

Methods

The assembly-free, K-mer-based SNP detection method (afkSNP), is designed to detect isolated SNPs directly on the reads in the WGS data. An isolated SNP is defined as a single nucleotide that is different in two samples, whereas the flanking regions are the same. To localize a base in a sample, we do no longer rely on its mapped position with respect to a reference sequence, but use the flanking regions of a base to define its location. Given a word size k, a location in one sample is thus said to be the same as a location in another sample if the (k-1)/2 bases before both locations, *and* also the (k-1)/2 bases after both locations are identical. If the base at the considered location (that is, the middle base of the word with size k) is different, the two samples are said to have a SNP difference. This method is covered by two steps:

1. For each sample, the complete list of words with length k is calculated. To limit sequencing errors, we filter on base quality scores, coverage, strand specificity and middle base ambiguity.
2. For each pair of samples, the word lists are compared by counting the number of words that have the same start and end, but have a different middle base. This yields a pairwise distance matrix that can be used for further analysis.

Once the word list of all samples is calculated, the comparison of samples is fast, does not require large amounts of memory and is highly parallelizable.

Results & Discussion

We have used the afkSNP method on two *Mycobacterium* data sets. First, we used 22 WGSs of identical *Mycobacterium bovis* (MBO) strains to investigate the stability and the reproducibility of the method. Next, we used a well-known set of 26 outbreak-related *Mycobacterium tuberculosis* (MTBC) samples to show the usefulness of the afkSNP method to rapidly suggest epidemiological connections between the samples. The data we used, has been generated in the context of the Patho-NGen-Trace project. The MTBC strains were provided by the National Reference Center for Mycobacteria (Borstel, Germany), the sequencing was performed by GenoScreen (Lille, France).

In contrast to reference-based methods used to detect SNPs, the afkSNP method, in a pairwise manner, compares the complete genomic content of all WGS samples, and not just what is common with the reference sequence. The use of only *isolated* SNPs does restrict the analysis, but at the same time this approach avoids clusters of SNPS that have been introduced by a single evolutionary event, thus providing a less disturbed counting of evolutionary events. Therefore, the afkSNP method covers a middle ground between reference-based SNPs and whole genome multi-locus sequence typing[1].

References

1. Jolley & Maiden 2010, BMC Bioinformatics 11:595

COMPARATIVE ANALYSIS OF BIOME-SPECIFIC MICROBIAL ASSOCIATION NETWORKS

Karoline Faust[1,2,*] & Jeroen Raes[1,2].
Dept. of Applied Biological Sciences[1], Vrije Universiteit Brussel, Dept. of Structural Biology, VIB[2].
[*]karoline.faust@vib-vub.be

Modern sequencing technology has made it possible to determine the composition of a microbial community and the relative abundances of its members. Community profiling experiments have now been carried out for thousands of samples coming from a variety of ecosystems, ranging from the ocean to the human gut. This large sample number opens the way to a comparative analysis of microbial communities in different biomes. While several such studies have been already conducted, none of them has yet considered the network level. Here, we present the results of a comparative analysis of biome-specific microbial association networks.

INTRODUCTION

The QIIME database[1] stores a large collection of uniformly processed 16S data from a variety of biomes. We have employed this rich data source to construct networks for 4 different soil biomes (tundra, grasslands, moist tropical forests and coniferous tropical forests) and 3 different host-associated biomes (gut, skin and oral cavity). We then compared various abundance matrix and network properties for these biomes.

METHODS

To construct the biome-specific networks, we applied our in-house ensemble-based network inference tool CoNet[2]. CoNet allows combining several similarity measures (we selected Pearson, Spearman, Kullback-Leibler dissimilarity and Bray Curtis dissimilarity) on the p-value level. P-values were computed for each method and each potential edge from both a permutation and a bootstrap distribution. To mitigate compositionality bias for Spearman and Pearson correlation, we renormalized each permuted vector pair with the full input matrix (ReBoot[2]). We then merged measure-specific one-sided p-values using Brown's method[3] and corrected merged p-values for multiple testing using Benjamini-Hochberg's procedure.

We computed alpha diversity using Shannon's index, richness with the Chao1 estimator (implemented in the R package vegan) and evenness using Sheldon's index, which in contrast to the more widely applied Pilou index is not biased by species number.

RESULTS & DISCUSSION

Closer inspection of the 7 networks showed that they reproduced known microbial relationships. For instance, the gut network partly replicates the co-occurrence networks defining the enterotypes[4], whereas the tundra network featured two clusters, one correlated and the other anti-correlated with pH.

When comparing network properties (Figure 1), we found that the soil networks had a significantly lower percentage of positive edges than the host networks. This finding was confirmed by body-area-specific networks constructed from QIIME-processed HMP[5] 16S data (which were not included in the data used for biome-specific network construction). After excluding sample number and sequencing depth as possible explanations, we computed the alpha diversity, richness and evenness of biomes and found that evenness was significantly anti-correlated with positive edge percentage. The higher richness, evenness and diversity of soil biomes as compared to host-associated biomes is in agreement with previous findings[6]. However, the high anti-correlation between evenness and positive edge percentage suggests that evenness plays an hitherto underappreciated role in shaping association networks. A possible reason for the impact of evenness could be the stronger sampling bias against rare taxa in uneven biomes, which might increase the positive edge percentage if negative relationships occur predominantly between rare taxa. Another reason might involve niche structure, which is easier to sample in uneven biomes, where only a few niches are within detection limits.

This study demonstrates that count matrix properties such as evenness and richness affect the properties of the resulting networks in unexpected ways and should be taken into account when interpreting microbial association networks.

Biome	Processed data OTU and sample number	Alpha diversity (Shannon index)	Richness (Chao 1)	Evenness (Sheldon index)	Number of nodes and edges	Positive edge percentage	Average cluster coefficient	Average path length	R2 of linear regression on node degree distribution
Coniferous forests	374 90	4.5	2.35	0.58	20 12	0	0	1.2	0.96
Moist forests	427 87	4.54	2.36	0.56	207 985	9.6	0.07	2.45	0.84
Grasslands	423 36	4.63	2.4	0.6	143 137	1.5	0	1.69	0.94
Tundra	348 33	3.97	2.18	0.41	291 1,954	39.6	0.38	2.41	0.78
Intestine	314 842	2.84	1.88	0.13	221 2,097	54.1	0.42	2.35	0.65
Oral cavity	334 408	2.91	1.79	0.12	218 755	74	0.41	2.44	0.76
Skin	929 1,323	3.47	2.23	0.12	565 2,736	78.7	0.36	2.92	0.85

***FIGURE 1**. Comparison of selected biome properties on the matrix and network level. Alpha diversity, evenness and richness were computed on count matrices rarefied to the same total count sum per sample.*

REFERENCES

1. Caporaso JG et al. *Nat. Methods* **7**, 335-336 (2010).
2. Faust K*, Sathirapongsasuti F* et al. *PLoS Comput. Biol.* **8**, e1002687 (2012).
3. Brown MB. *Biometrics* **31**, 987-992 (1975).
4. Arumugam M*, Raes J* et al. *Nature* **473**, 174-180 (2011).
5. Methé BA et al. Nature **486**, 215-221 (2012).
6. Fierer N & Lennon JT. *Am. J. Botany* **98** (3), 439-448 (2011).

TOWARDS A SOFTWARE-INDEPENDENT TAXONOMIC PROFILING METHOD OF MICROBIAL METAGENOMES

Koen Illeghems, Zoi Papalexandratou, Luc De Vuyst & Stefan Weckx[*]
Research Group of Industrial Microbiology and Food Biotechnology (IMDO), Faculty of Sciences and Bioengineering Sciences, Vrije Universiteit Brussel, Brussels, Belgium; [*]stefan.weckx@vub.ac.be

An essential question in the metagenomic analysis of a microbial ecosystem is the identification of the source microorganism of each read or contig, an aspect that is still in its infancy. Moreover, the result is often dependent on the analysis software used. Yet, to perform software-independent taxonomic profiling, several computational methods are available, tackling either a composition-based or a similarity-based approach. It is unclear which of these methods result in the best estimate of the microbial species diversity of an ecosystem. Indeed, similarity-based methods rely on the presence of a close evolutionary relative in the database used and are known to be computationally expensive. In the case of (supervised) composition-based methods, it is (often incorrectly) assumed that the genomes available in public databases are representative for the microorganisms present in the targeted ecosystem. In this study, the cocoa bean fermentation ecosystem was used as a case-study to investigate the microbial species diversity using similarity-based and composition-based taxonomic profiling methods. Validation of the results showed that a combination of methods provides an optimal overview of the members present in the ecosystem.

INTRODUCTION

The microbial species diversity of spontaneous cocoa bean fermentation processes has been investigated through the application of culture-dependent and culture-independent techniques[1-4]. This has resulted in a good knowledge of this peculiar microbial ecosystem, which is dominated by species such as *Hanseniaspora* sp., *Saccharomyces cerevisiae*, *Lactobacillus fermentum*, *Lactobacillus plantarum*, and *Acetobacter pasteurianus*. However, it is known that both approaches have some drawbacks, undermining an accurate view on the microbial composition of this ecosystem, hence implying that more, yet unidentified species might play a role in the fermentation process. The present study investigated the microbial communities of a single sample of a spontaneous cocoa bean box fermentation process by applying a combination of taxonomic profiling methods on metagenomic sequence data.

METHODS

To investigate the performance of different similarity-based and composition-based approaches, when applied on a real metagenomic data set, the microbial communities of a sample of a spontaneous cocoa bean box fermentation process were analyzed. This was achieved by 454 pyrosequencing followed by taxonomic profiling of metagenomic reads using both similarity-based and composition-based computational methods (CARMA[5], MEGAN[6], MetaPhyler[7], PhymmBL[8], RAIphy[9] and SmashCommunity[10]). Furthermore, to assess the metabolic potential of the members of this ecosystem, a functional analysis was performed. To obtain the best assembly, different assemblers were applied, including genomic assemblers (Newbler[11], CABOG[12], CAMERA[13]) as well as a metagenomic assembler (Genovo[14]). Annotation of the assembled data was carried out by the GenDB[15] platform.

RESULTS & DISCUSSION

Operational taxonomic units that were consistently predicted by the different taxonomic profiling tools were taken into account to avoid a software-dependent outcome. This approach identified both prevailing microbial species (*Hanseniaspora uvarum, Hanseniaspora opuntiae, Saccharomyces cerevisiae, Lactobacillus fermentum,* and *Acetobacter pasteurianus*) as well as rare microbial species (*Erwinia tasmaniensis*, *Lactobacillus brevis*, *Lactobacillus casei*, *Lactobacillus rhamnosus, Lactococcus lactis*, *Leuconostoc mesenteroides*, and *Oenococcus oeni*). Furthermore, the restricted viral diversity, dominated by *Myoviridae* and *Siphoviridae*, reflected *Lactobacillus* as the dominant host. These results were subsequently compared with previously obtained culture-dependent and culture-independent data. Overall, the presence and abundance of the dominating members of this ecosystem were comparable, demonstrating the validity of the software-independent taxonomic profiling. In addition, a wider community diversity was retrieved by the metagenomic sequencing approach compared with the other approaches, situated mainly within the *γ-Proteobacteria* and the fungal members. Hence, a complete and more reliable insight into the community diversity of the cocoa bean fermentation sample studied could be obtained. This indicates the superiority of metagenomic sequencing using a combination of similarity-based and composition-based taxonomic profiling methods. Concerning a functional analysis of the data, the metagenomic assembler Genovo provided the best assembly. Subsequent reconstruction of metabolic pathways revealed new insights into the cocoa bean fermentation ecosystem, such as citrate metabolism and pectin degradation.

REFERENCES

1. Ardhana MM *et al. Int J Food Microbiol* **86**, 87-99 (2003).
2. Camu N *et al. Appl Environ Microbiol* **73**, 1809-1824 (2007).
3. Nielsen DS *et al. Yeast* **22**, 271-284 (2005).
4. Papalexandratou Z *et al. Food Microbiol* **28**, 964-973 (2011).
5. Gerlach W & Stoye J *NAR* **39**, e91 (2011).
6. Huson DH *et al. Genome Res* **17**, 377-386 (2007).
7. Liu B *et al. BMC Genomics* **12**, S4 (2011).
8. Brady A *et al. Nat Methods* **6**, 673-676 (2009).
9. Nalbantoglu OU *et al. BMC Bioinformatics* **12**, 41 (2011).
10. Arumugam M *et al. Bioinformatics* **26**, 2977-2978 (2011).
11. Margulies M *et al. Nature* **437,** 376-380 (2005).
12. Myers EW *et al. Science* **287**, 2196-2204 (2000).
13. Sun S *et al. NAR* **39,** D546-D551 (2011).
14. Laserson J *et al. J Comput Biol* **18**, 429-43 (2011).
15. Meyer F *et al. NAS* **31**, 2187-2195 (2003).

IDENTIFYING INTERACTION PATTERNS IN HUMAN MICROBIOTIA

X. Wang[1], D. Bogaert[1], W.T. Hendriksen[1], G. Biesbroek[1], E.A.M. Sanders[1], K. Trzcinski[1], J. Wallinga[2], M.J.C. Eijkemans [3].*

[1]Department of Pediatric Immunology and Infectious Diseases, UMC Utrecht, Utrecht, The Netherlands. [2]Department of Infectious Diseases Epidemiology, National Institute of Public Health and the Environment, Bilthoven, The Netherlands. [3]Julius Center for Health Sciences and Primary Care, UMC Utrecht, Utrecht, The Netherlands

The human microbiome interacts with the human host to prime immunity and to maintain host health. High-throughput sequencing methods enable us to investigate the microbial composition of ecological niches. These methods generate per sample the relative or absolute abundance of a large number of microbial species. To identify symbiotic and competitive interaction patterns between microbial species, we need statistical clustering (or: unsupervised) methods. The optimal method will depend on the particular patterns of interest (symbiotic or competitive) and may differ between datasets.

METHODS

For a given metagenomic dataset with given patterns of interest, we propose several clustering approaches to determine the optimal combination of number of clusters, distance function, and linkage method by calculating the silhouette index. We also proposed a new approach to detect the completion pattern based on the clique identification. The proposed strategies and methods were applied to a published dataset containing 16S-rDNA-based sequencing data of 96 nasopharyngeal samples obtained from children 18 months of age.

RESULTS

For the tested dataset, the combination of Pearson correlation distance, average linkage and a number of 21 clusters performed best to identify symbiotic patterns, while a competitive pattern also obtained with Pearson correlation distance and average linkage.

CONCLUSION

With this strategy, we have designed a reliable tool to identify bacterial correlations within a given metagenomic dataset. The value of this approach was illustrated by the observation of several clinically relevant bacterial interactions in our 16S-rDNA sequence dataset. This flexible approach can be applied to any microbiome data sets for identifying (symbiotic or competitive) interaction patterns

COMPARATIVE METAGENOMICS BY CROSS-ASSEMBLY

Bas E. Dutilh[1,2,*].

*Centre for Molecular and Biomolecular Informatics, Nijmegen Centre for Molecular Life Sciences, Radboud University Medical Centre, Nijmegen, The Netherlands[1]; and Department of Marine Biology, Institute of Biology, Federal University of Rio de Janeiro, Rio de Janeiro, Brazil[2]. *bedutilh@gmail.com*

Cross-assembly of different metagenomes is a fast and insightful way to obtain information about sequences that are shared between the samples, represented by cross-contigs. Importantly, cross-assembly is independent of an annotated reference database, providing a way to also handle unknown sequences. The cross-assembly tool crAss allows a rapid analysis of these cross-contigs. First, it provides cross-contig-based similarity scores between all metagenome pairs. Second, crAss creates insightful images displaying the inter-relationships between samples. Third, it generates occurrence profiles of the cross-contig sequences across metagenomes that can be used to discover related sequences, aiding further assembly and interpretation.

INTRODUCTION

Determining the interrelationships between metagenomes from different biomes or different time points is important to understand the microbial world around us. Mapping metagenomic sequences to a reference database of known genes is a feasible approach to transfer taxonomical and functional annotations to sequence reads. However, it can limit the amount of data that can be analyzed because the majority of the sequencing reads in difficult-to-annotate datasets, such as viral metagenomes from biomes other than the human microbiome, lack known homologs[1]. A promising alternative is reference-independent comparative metagenomics by cross-assembly[2].

METHODS

First, a cross-assembly should be created. This can be done by combining the metagenomic datasets and assembling all the sequencing reads with a (metagenome) assembly tool of choice. Next, crAss takes as input the assembly results file (ACE format) as well as one FASTA or FASTQ file per metagenomic dataset, which is necessary to allow the program to identify from which dataset every read originated. The output of crAss consists of several, easily parsable files including a file that indicates how many reads from each metagenome were assembled into each cross-contig; distance matrices between all metagenomes using four distance formulas; and PNG image files displaying these similarities (Figures 1 and 2).

RESULTS & DISCUSSION

First, I will show that crAss identifies meaningful similarities between metagenomes based on shared sequences, by using simulated and real data. Next, I will address the issue of chimerization: combining sequences from different genomes into a single cross-contig. I will discuss the problems that may arise from chimerization, and suggest ways to assess their impact. Finally, I will show that the occurrence profiles of cross-contigs contain valuable information for further assembly and interpretation of the resulting sequences.

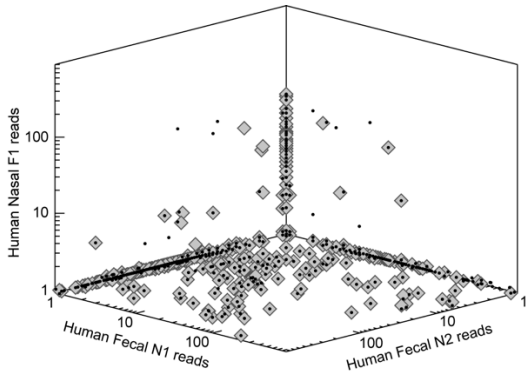


FIGURE 1. Cross-assembled metagenomic reads from one human nasal sample and two human fecal samples. Each gray diamond represents a contig. The X, Y and Z coordinates indicate the number of incorporated reads from the metagenomes mentioned along the axes. Image generated by crAss[2].

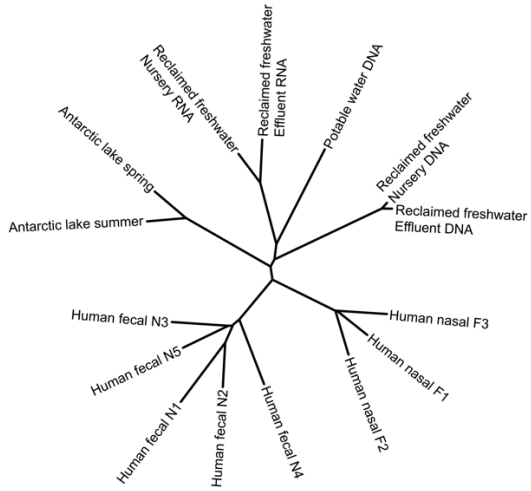


Figure 2. Cladogram representing the distance between metagenomes based on the fraction of cross-assembled contigs between all sample pairs. Image generated by crAss[2].

REFERENCES

1. Mokili J *et al*. *Curr Opin Virol* **2**, 1-15 (2013).
2. Dutilh BE *et al*.. *Bioinformatics* **28**, 3225-3231 (2012).

ORCAE: ONLINE RESOURCE FOR COMMUNITY ANNOTATION OF EUKARYOTES

Lieven Sterck[1,2,*], *Kenny Billiau*[1,2], *Thomas Abeel*[1,2], *Pierre Rouzé*[1,2] & *Yves Van de Peer*[1,2].
[1] *Department of Plant Systems Biology, VIB, Technologiepark 927, 9052 Ghent, Belgium.*
[2] *Department of Plant Biotechnology and Bioinformatics, Ghent University, Technologiepark 927, 9052 Ghent, Belgium.*
[*] *lieven.sterck@psb.vib-ugent.be*

Conducting gene and genome annotation typically relies on diverse information resources going from sequence (e.g. the genomic sequence, transcript and protein alignments) to expression data (e.g. microarray or read count analysis) depending on whether structural or functional annotation is performed. To help researchers doing gene annotation while having access to these different data types, we developed ORCAE (Online Resource for Community Annotation of Eukaryotes), a web-technology-compliant portal for use in community genome annotation efforts.

INTRODUCTION

ORCAE allows browsing and on the fly editing of gene descriptions as well as gene structures, moreover all manual curations are immediately visible for other users. The portal will store all the modification from annotators in the database so for each locus a history of modifications is available.

Through its interface, ORCAE offers easy access to precomputed information that greatly facilitates the work of a curator. The gene page offers several informative graphics with a focus on the quality of the gene structure (e.g. Multiple alignments of similar proteins, tiling array information …) helping the human annotators in improving the proposed automated annotation. Annotators can use this system with the build-in GenomeView interface to check/modify gene structures. A unique feature from ORCAE is that the portal is highly dynamic, unlike systems that have a more static implementation. On modification of a gene model, all the available information (eg. protein similarity, transcript alignments, etc.) is immediately updated and presented on the gene page.

ORCAE can both be used to coordinate annotation efforts in the course of the project as well as to present published genomes to the public. It has therefore 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).

Currently it offers public access to 12 eukaryotic genome projects and restricted access to another 9 genomes.

ORCAE is available at :
http://bioinformatics.psb.ugent.be/orcae/.

REFERENCES

1. Sterck L *et al*. *Nat. Methods* ***9****, 1041* (2012).

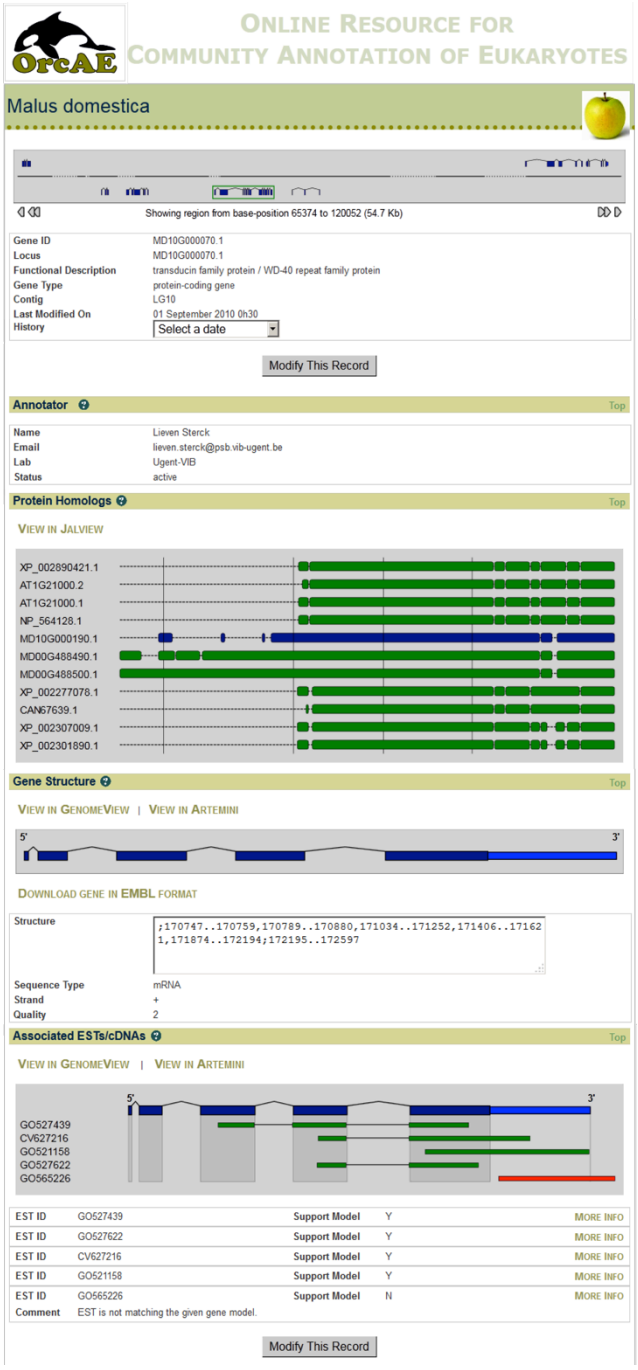


FIGURE 1*. Gene page in the ORCAE resource. Through the extensive use of graphical representations, a clear overview of the data is provided to assist users to easily asses the quality and correctness of the offered annotations for a given gene locus.*

COMPARING FRAGMENTATION SPECTRA FROM TWO PARASITIC WORM SPECIES TO DISCOVER UNIQUE PEPTIDES

Şule Yılmaz[1,2], Bjorn Victor[3], Niels Hulstaert [1,2], Giulia Gonnelli [1,2], Pierre Dorny[3], Magnus Palmblad[4], Lennart Martens[1,2].

Dept. of Medical Protein Research, VIB, Ghent, Belgium[1]; Dept. of Biochemistry, VIB, Ghent, Belgium[2]; Veterinary Helminthology Unit, Dept. of Biomedical Sciences, Institute of Tropical Medicine, Antwerp, Belgium[3]; Biomolecular Mass Spectrometry Unit, Dept. of Parasitology, Leiden University Medical Center, Leiden, The Netherlands[4]

*lennart.martens@ugent.be

In mass spectrometry based proteomics, spectra are typically assigned to peptides by database searches. Even though complete genomes, and therefore derived proteome databases, are available for many species, most species have limited or no information at all. *Taenia solium* and *Taenia hydatigena* are two parasitic worm species with only a limited protein database. Current diagnostics tests fail to distinguish these two species, one of which is an important cause of disease in humans. This study therefore investigates primarily the spectra to determine unique spectra in either of the two datasets.

INTRODUCTION

Proteomics enables the identification of proteins, and the derived field of proteogenomics provides an opportunity to assign identified peptides to a genome. Although sequenced genomes are now available for many species, genome annotation is lagging behind. *Taenia solium* and *Taenia hydatigena* are closely related tapeworm species where genome information is limited. The larval stages of both these species can infect pigs, while humans can only be infected by *T. solium*. When investigating the prevalence of *T. solium* in pigs, a diagnostic test is needed to distinguish between *Taenia* species and assess the risk to humans. However, current tests cannot make this species-level distinction. A possible solution would be (i) to identify species-specific proteins, and (ii) to focus a diagnostic test on these. To achieve this, we compare spectra between datasets from each species to determine unique spectra. Full proteogenomics searches are then used as a follow-up for interesting leads.

METHODS

The strategy can be divided into four steps: preprocessing, quality based classification and filtering, similarity calculation, and candidate visualization (Figure 1). The datasets come from the two taenia species. The first step eliminates any spectra that are assigned to host proteins in addition to common contaminants at FDR=1%. That is followed by filtering based on their quality[1]. Then, similarities between the remaining spectra are calculated and unique spectra are retrieved. After that, unique spectra are matched to the draft *T. solium* protein database and a six reading frame (6RF) translation search is performed[2,3].

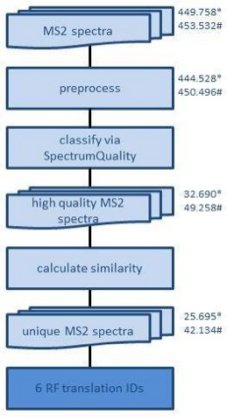


***FIGURE 1**. : Workflow to discover unique spectra. Numbers indicate dataset sizes. *T. hydatigena #T. solium*

RESULTS & DISCUSSION

Five different similarity scores (dot product, MSE, MdSE, Pearson's and Spearman's correlation coefficient) are calculated to compare *T. hydatigena* spectra against *T. solium* ones. A Java tool is developed to visually inspect the similarity between two spectra in light of the obtained similarity scores. After analysis of the different scores, Spearman's correlation coefficient at threshold=0.6 is applied to noise-filtered, log-scaled intensities to determine unique spectra (Figure 2).

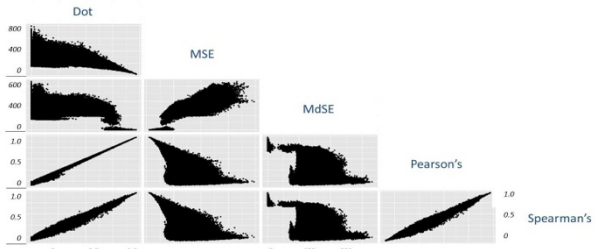


***FIGURE 2**. Scatterplots for different similarity scores across all spectra*

Figure 3 shows how unique spectra are assigned to the *T.solium*-6RF database at FDR=1%. Some peptides are spread over different regions of the gel, indicating isoforms or modification states.

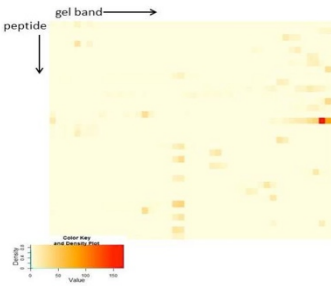


***FIGURE 3**. Assigned peptides by 6RF db search for unique spectra at T.solium*

REFERENCES

1. Flikka, K *et al.Proteomics* **6(7)**, 2086–94 (2006)
2. Vaudel, M *et al Proteomics* **11**, 996-999 (2011)
3. http://peptide-shaker.googlecode.com

IDENTIFYING LOSSES AND EXPANSIONS OF SELECTED GENES FAMILIES IN INCOMPLETE GENOMIC DATASETS

Arnaud Di Franco[1,2 *], *Mark Hanikenne*[2,3], *Denis Baurain*[1,2].

[1]*Eukaryotic Phylogenomics, Department of Life Sciences, University of Liège, B-4000 Liège, Belgium ;*[2]*PhytoSYSTEMS, University of Liège, B-4000 Liège, Belgium ;* [3]*Functional Genomics and Plant Molecular Imaging, Center for Protein Engineering (CIP), Department of Life Sciences, University of Liège, B-4000 Liège, Belgium;*
**arnaud.difranco@gmail.com*

Plantae (Archaeplastida) are a natural group of organisms with plastids of primary endosymbiotic origin. Within this group, members of the red algae show evidence of a reduction of their genomic content. In this work, we designed a bioinformatics approach to investigate the few, sometimes incomplete, genomic datasets available for red algae, with the purpose of pointing out possible gene family losses and expansions. Our pipeline first populates a relational database with precomputed orthology relationships between green plant genomes and red algal datasets and then efficiently queries the database for computing statistics of losses and expansions for a series of gene families of interest.

INTRODUCTION

Primary plastids have been acquired by eukaryotes through a unique event of endosymbiosis with a cyanobacteria. This ancestral photosynthetic eukaryote gave rise to three monophyletic groups, glaucophytes, green plants and red algae. The latter have generally smaller genomes with less coding genes compared to the two others (especially green plants). We suppose that those losses are due to specific environmental conditions having affected the common ancestor of red algae. To prove our hypothesis, we developed a pipeline to examine presence/absence of gene families of interest in this group for which only few genomes are available.

METHODS

We built an inventory containing the genes of interest by searching UniprotKB with specific keywords. Each gene was assigned to one or more categories depending on its functional annotations. Meanwhile, we collected the genomes of several green plants and defined orthologous groups (OGs). To consider only the gene families that were present in the common ancestor of Plantae (POGs), we used a taxonomic filter ensuring a minimal representation of each child lineage. Then, we used our inventory to partition the POGs into annotated and anonymous gene families. To identify the genes of interest in red algae, we built Hidden Markov Model profiles (pHMMs) from the aligned POGs and searched for each pHMM in the red algal genomic datasets. Finally, we used a BLAST Best Reciprocal Hit criterion (BRH) to discard matches to paralogous genes, in which, to be retained as orthologous, each red algal hit had to match back to a green plant sequence belonging to the pHMM.

Orthology relationships were loaded into the database, along with details about the inventory, functional categories and pHMMs properties. At first, we queried the database to retrieve the groups of pHMMs corresponding to the various categories of our inventory. These real groups (RGs) were used to look for orthologous genes into red algal datasets. To study losses, counts of red algal orthologs were converted to boolean values (presence/absence), whereas all orthologs were counted individually to study family expansions. Then, to verify whether the observed losses/expansions were significant, we compared the counts obtained for each category to background distributions generated from 1000 control groups of pHMMs (CGs). CG-pHMMs were selected so as to match the properties of the RG-pHMMs. Hence, for each pHMM properties, we performed a Kolmogorov-Smirnov test (KS) of its distribution in each of our RGs versus all of our pHMMs. To select the properties that we had to take into account when assembling the CGs, we sorted them by the geometric mean of their p-values over the different RGs. This allowed us to retain the seven most critical properties. Each CG was assembled as follows: (1) a number of pHMMs equal to the size of the RG are picked at random, (2) a KS is carried out for each of the seven properties between the RG and the candidate CG, (3) if the geometric mean of the seven KS p-values is greater than 0.05, the CG is accepted; otherwise the algorithm goes back to step 1.

RESULTS & DISCUSSION

Losses/expansions were investigated for each of our functional categories in five genomes of red algae. While gene losses were difficult to confidently identify, some interesting family expansions were detected (Figure 1). Further, we checked that the approach also worked on transcriptomic data.

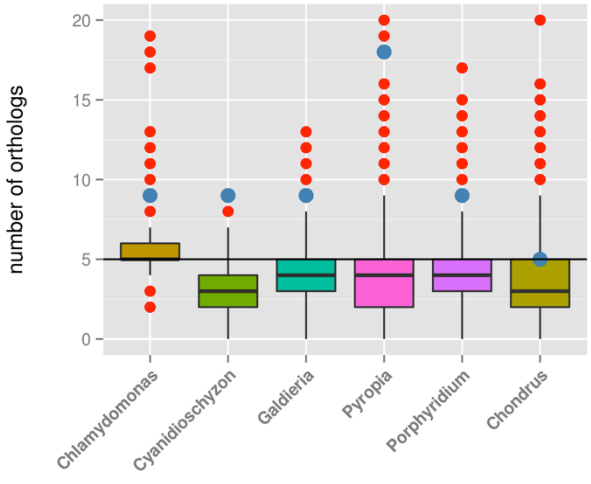


***FIGURE 1**. Expansions in iron transport-related gene families in five species of red algae.*

UNRAVELLING THE GENETIC BASIS OF *FUSARIUM* SUGARBEET WILT DISEASE

Ronnie de Jonge[1,2*], Klaas Vandepoele[1,2], Yves Van de Peer[1,2] & Melvin D. Bolton[3].
*Dept. of Plant Systems Biology, VIB, Ghent, Belgium[1] ; Dept. of Plant Biotechnology and Bioinformatics, Ghent University, Ghent, Belgium[2] ; Northern Crop Science Laboratory, United States Department of Agriculture – Agricultural Research Service, Fargo, ND, United States[3]. *ronnie.dejonge@psb.ugent.be*

Sugarbeet is an important source of sucrose throughout the world. Although yield and quality losses caused by *Fusarium* spp. impart significant economic losses for sugarbeet growers, little is known about the molecular genetics of pathogenicity of these species. Recently, a new disease was reported in Minnesota that caused Fusarium yellows-like symptoms, but was shown to be more aggressive. Preliminary phylogenetic analyses suggest that a novel *Fusarium* sp., tentatively called *F. secorum*, is responsible for this disease. To gain insights into the mechanisms that confer the organisms' pathogenicity and investigate the genetic background of this disease we sequenced the genomes of *F. secorum*, *F. oxysporum* f. sp. *betae* and a related non-pathogen *F. acutatum* and compared them to each other, and to published *Fusarium* genomes.

INTRODUCTION

Sugarbeet is an important source of sucrose for natural sweetening throughout the world. The sugarbeet growing areas of North Dakota and Minnesota constitute the largest sugarbeet production area in the US. Yield and quality losses caused by *Fusarium* spp. impart significant economic losses for sugarbeet growers. Fusarium yellows, caused by *Fusarium oxysporum* f. sp. *betae*, is a common sugarbeet disease in the US. The disease is characterized by interveinal chlorosis, wilting of foliage, and vascular discoloration of the taproot, often leading to plant death (Figure 1). Recently, a new disease was reported in Minnesota that caused Fusarium yellows-like symptoms, but was shown to be more aggressive. Disease symptoms resembled those of Fusarium yellows, but differed by discoloration of petiole vascular elements, seedling infection and more rapid death of plants. In addition, the *Fusarium* species causing the novel disease could be isolated from petioles, unlike other *Fusarium* sugarbeet pathogens. Sequence analysis of *ELF1α* did not significantly match any known *Fusarium* species, suggesting that a novel *Fusarium* sp. is responsible for this disease. As such, this disease has been tentatively named Fusarium yellowing decline in order to differentiate from Fusarium yellows and the disease agent *F. secorum*.

Preliminary phylogenetic analyses revealed a high similarity of *F. secorum ELF1α*, *Calmodulin* and *mtSSU* to *F. acutatum*, a *Fusarium* sp. that is non-pathogenic on sugarbeet and not found in the US, whereas similar analyses using the *beta-tubulin* gene pointed towards relatedness to *F. oxysporum* f. sp. *betae.*

RESULTS & DISCUSSION

To investigate the genetic background of the recently discovered sugarbeet pathogen *F. secorum* and simultaneously assess genetic factors involved in sugarbeet pathogenicity, we have set out to determine its genome sequence as well as that of *F. oxysporum* f. sp. *betae* and *F. acutatum*.

Using Illumina sequencing we generated short-insert (500 bp) and long-insert (5 Kbp) sequence libraries that were used for assembly.

***FIGURE 1**. Fusarium yellows symptoms on sugarbeet*

The assembled genomes range in size between 44 Mbp (*F. acutatum*), 49 Mbp (*F. secorum*) and 53 Mbp (*F. oxysporum* f. sp. *betae*). We then combined *ab initio* gene predictions with protein alignments[1] resulting in the prediction of 14,629; 15,872; 15,804 and 16,670 genes for *F. acutatum*, *F. secorum*, *F. oxysporum* f. sp. *betae* strain-1 and strain-2 respectively; numbers comparable to those of *F. graminearum* (13,321), *F. verticillioides* (14,169) and *F. oxysporum* f. sp. *lycopersici* (17,708). To assess evolutionary relationships across all taxa, we established a set of 7,847 core ortholog groups with 1 to 1 relation in all mentioned taxa by incorporating all genome data in PLAZA[2] and subsequently used one thousand randomly selected ortholog groups to perform phylogenetic analyses. These analyses suggest that *F. secorum* is most closely related to one *F. oxysporum* f. sp. *betae* strain whereas surprisingly the other *F. oxysporum* f. sp. *betae* strain clusters with *F. oxysporum* f. sp. *lycopersici*.

REFERENCES

1. Haas B. et al. *Genome Biol*, 2008, 9:R7.
2. Proost S. et al. *Plant Cell*, 2009, 21:3718.

SEQUENCE BASED GENOTYPING: APPLICATIONS FOR PLANT BREEDING

Erwin Datema[*], *René Hogers, Nathalie J. van Orsouw, Michiel J.T. van Eijk and Antoine Janssen.*
Keygene N.V., Agro Business Park 90, P.O. Box 216, 6700 AE Wageningen, The Netherlands. [*]*eda@keygene.com*

We present Sequence Based Genotyping (SBG), a technology which combines the genome complexity reduction of AFLP® with the high throughput sequencing capacity of the Illumina® MiSeq™ and HiSeq™ platforms to score hundreds to thousands of genetic markers distributed randomly across a genome. We highlight the general applicability of SBG in plant breeding through several examples.

INTRODUCTION

SBG is a cost-efficient and highly robust SNP discovery and genotyping method that does not require prior genome sequence information and performs well on a wide range of genome sizes and ploidy levels. Example applications include screening large collections of germplasm for novel variation, and building high-density, accurate genetic maps from mapping populations. Through incorporation of selective nucleotides during sequencing template construction, the user can strike an optimal balance between the desired number of genotyped SNPs and the level of sample multiplexing during sequencing.

METHODS

To analyze the data, we have created a highly automated, parallel analysis workflow that can be executed either from the Linux command line or through a Galaxy web interface. The software can either use an existing genome sequence or construct a complexity-reduced reference sequence de novo from the SBG sequence data. The workflow exploits both widely used bioinformatics tools such as BWA and GATK to perform genotyping, and novel tools for removal of repetitive sequences and haplotyping. The analysis software runs on commodity hardware and scales well with increasing data volumes. Additionally, it features customized filtering and postprocessing options tailored towards the analysis of both germplasm and breeding populations. Outputs include machine-parsable VCF and LOC file formats, as well as a human-readable tabular format.

RESULTS & DISCUSSION

We have applied SBG to a variety of crop plants including leek (*Allium ampeloprasum*), rapeseed (*Brassica napus*), cabbage (*Brassica oleracea*), pepper (*Capsicum annuum*), cucumber (*Cucumis sativus*), cotton (*Gossypium hirsutum*), lettuce (*Lactuca sativa*), rice (*Oryza sativa*), tomato (*Solanum lycopersicum*), spinach (*Spinacia oleracea*), and wheat (*Triticum aestivum*). We will present a selection of these data to highlight relevant applications of SBG in plant breeding, such as QTL mapping, Bulk Segregant Analysis and Genome Wide Association Mapping. To facilitate straightforward exploitation of this technology, we have created an "SBG 100-Kit" that includes optimized AFLP primers and is available for research purposes.

CONVERGENT GENE LOSS FOLLOWING GENE AND GENOME DUPLICATIONS CREATES SINGLE-COPY FAMILIES IN FLOWERING PLANTS

Riet De Smet[1,2], *Keith Adams*[1,3], *Klaas Vandepoele*[1,2], *Steven Maere*[1,2] & *Yves Van de Peer*[1,2,*].
Department of Plant Systems Biology, VIB, Technologiepark 927, 9052 Ghent, Belgium [1]; *Department of Plant Biotechnology and Bioinformatics, Ghent University, Technologiepark 927, 9052 Ghent, Belgium* [2] ; *Department of Botany, University of British Columbia, 6270 University Blvd, Vancouver, BC, V6T 1Z4, Canada*[3].
[*]*yves.vandepeer@psb.vib-ugent.be*

The importance of gene gain through duplication has been appreciated for a long time. Contrary, the importance of gene loss has only recently attracted attention. Indeed, studies in organisms ranging from plants to worms and humans suggest that duplication of some genes might be better tolerated than that of others. Here we have undertaken a large-scale study to investigate the existence of duplication-resistant genes in the sequenced genomes of 20 flowering plants. We demonstrate that there is a large set of genes that is convergently restored to single-copy status following multiple genome-wide and smaller-scale duplication events. We rule out the possibility that such a pattern could be explained by random gene loss only and therefore propose that there is selection pressure to preserve such genes as singletons. This is further substantiated by the observation that angiosperm single-copy genes do not comprise a random fraction of the genome, but instead are often involved in essential housekeeping functions that are highly conserved across all eukaryotes. Furthermore, single-copy genes are generally expressed more highly and in more tissues than non-single-copy genes, and they exhibit higher sequence conservation. Finally, we propose different hypotheses to explain their resistance against duplication.

INTRODUCTION

Following Ohno[1], gene duplication has been repeatedly reported to play an important role in evolution. For instance, mechanisms such as sub- or neofunctionalization underlie the evolution of many novel gene functions. Conversely, gene duplication can also be strongly deleterious and has been associated with diseases such as Parkinson[2] and cancer[3]. The full complement of genes for which duplication seems not to be tolerated, to the extreme that some genes occur as one single copy in any genome, is currently unknown. The availability of such a set of genes might, however, reveal whether general evolutionary and functional characteristics could explain the deleterious effects of their duplication. To assess selection against retention of certain gene duplicates and to study the relationship between gene duplicability and gene function, it is necessary to study a large number of genomes. The angiosperm lineage is especially well-suited to study duplication resistance because all angiosperm genes have repeatedly been subject to duplication due to the occurrence of multiple shared and independent whole-genome duplication events[4,5]. Here, we take advantage of the increasing number of available sequenced angiosperm genomes to identify duplication-resistant genes at high resolution.

METHODS

To identify genes that are single-copy in a large number of angiosperm genomes, we used the orthologous groups (OGs), predicted by the OrthoMCL method that are stored in the PLAZA 2.5 database[6]. These orthologous groups span 20 angiosperm genomes. Single-copy genes *sensu stricto* are defined as genes that are conserved in all angiosperm genomes and with a one-to-one orthology relationship in these genomes, i.e. they have remained single-copy since the angiosperm common ancestor or have consistently been restored to single-copy status following duplication. We developed a phylogenetic approach to validate the single-copy status of the obtained orthologous groups and to assess the possibility that paralogs were erroneously excluded from the OGs which would lead to false positive predictions of single-copy OGs.

RESULTS & DISCUSSION

By comparing 20 sequenced angiosperm genomes, we show that, despite the large number of small- and large-scale duplication events that have taken place, there exists a set of genes that has been repeatedly restored to single-copy status. Because the observed number of single-copy families greatly exceeds what can be expected from random gene loss effects, this suggests that selection promotes convergent evolution of these genes to single-copy status across angiosperms.

We found this set of genes not to be a random fraction of the genome but to encode housekeeping and other essential functions, as suggested by their functional enrichment, conservation throughout the eukaryotic tree, and expression breadth.

REFERENCES

1. Ohno S. Evolution by gene duplication. Springer (1970).
2. Singleton AB *et al. Science* **302,** 841 (2003).
3. Seeger R *et al. N Engl J Med* **313**,1111-1116 (1985).
4. Van de Peer Y *et al. Trends Plant Sci* **14**, 680-688 (2009).
5. Blanc G & Wolfe KH *Plant Cell* **16**, 1667-1678 (2004).
6. Van Bel M *et al. Plant Physiol* **158**, 590-600 (2012).

INTEGRATING GENE REGULATORY NETWORK INFERENCE SOLUTIONS FOR THE ABIOTIC STRESS RESPONSE IN *ARABIDOPSIS THALIANA*

Vanessa Vermeirssen[1,*], *Inge De Clercq*[1], *Frank Van Breusegem*[1] & *Yves Van de Peer*[1].
Department of Plant Systems Biology[1], *Department of Plant Biotechnology and Bioinformatics*[1], *VIB, Ghent University, Technologiepark 927, 9052 Gent,* [*]*vamei@psb.vib-ugent.be*

In order to elucidate the molecular mechanisms of the abiotic stress response in *Arabidopsis thaliana* at a systems level, we inferred gene regulatory networks through reverse engineering of a microarray gene expression compendium of 283 abiotic stress conditions. Through rank aggregation, we joined the prioritized regulatory interactions of the ensemble module networks algorithm LeMoNe, the mutual information direct network inference method CLR and the double two-way t-test method TwixTrix. Using in silico validation by reported protein-DNA and regulatory interactions and experimental validation by Nanostring nCounter analysis, we demonstrate that ensemble reverse-engineering generates robust biological hypotheses of regulatory interactions. Our combined computational and experimental approach identified key regulatory mechanisms through which plants deal with abiotic stress at a systems level.

INTRODUCTION

The plant responds to its changing environment by fine-tuned regulation through complex gene regulatory networks. Due to experimental challenges, only a fair number of gene regulatory interactions between transcription factors and their target genes have been experimentally mapped for *Arabidopsis*. However, these data can be predicted through reverse-engineering. Reverse-engineering infers gene regulatory networks based on the principle that the activity of transcription factors is embedded in their expression profiles. Benchmark studies in bacteria and yeast have shown that no single best reverse-engineering method exists: different methods show different biases in detecting regulatory relationships and act therefore complementary[1]. Therefore, the combination of the results of different network inference algorithms into one ensemble solution has recently been explored[1]. Until now, reverse-engineering has only been limitedly applied in Arabidopsis.

METHODS

We applied four different network inferences on a specific microarray compendium of 283 abiotic stress conditions: two different parameter settings of the stochastic Bayesian module network algorithm LeMoNe, the mutual information direct algorithm CLR and the double two-way t-test method TwixTrix. We constructed an abiotic stress gene regulatory network (GRN) of 200000 regulatory interactions from the ensemble solution obtained by rank aggregation of these four predictions. For in silico validation, we assembled a set of 50000 regulatory interactions from experimental protein-DNA interactions and transcription factor perturbed expression profiles. For experimental validation, we conducted Nanostring nCounter analysis[2] of gain and loss of function mutants for 7 transcription factors and 110 target genes under control and salt stress conditions.

RESULTS & DISCUSSION

In silico benchmark analysis showed that the ensemble solution has an as good performance as the individual solutions. By clustering the abiotic stress GRN in coregulated modules, we demonstrated the functional coherence of coregulated target genes by Biological Process Gene Ontology (GO) enrichment, Aracyc plant metabolic pathway enrichment, the presence of functional gene-gene links and experimental protein-protein interactions, the enrichment for known oxidative stress genes and the overrepresentation of cis-regulatory motifs. Example stress-related modules illustrated the power of the ensemble reverse-engineering in recovering known biological information, as well as providing novel hypotheses in the abiotic stress response. For the experimental benchmark set obtained through Nanostring nCounter analysis, we were able to confirm over half of the predicted interactions. Here, the ensemble rank aggregation clearly outperformed the individual reverse-engineering methods.

In conclusion, we showed that integrating solutions of network inference from abiotic stress gene expression profiles not only advances holistic understanding of the abiotic stress response and its key regulators, but also offers high potential as hypothesis generator for time- and cost-efficient design of experiments.

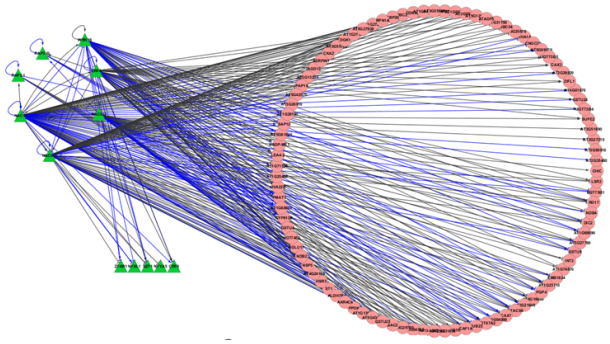


FIGURE 1*. The Nanostring nCounter experimental abiotic stress regulatory network. Blue edges = true positives, black edges = false negatives.*

REFERENCES

1. Marbach D. et al. *Nat Methods*. **9**, 796-804 (2012).
2. Geiss G. et al. *Nat Biotechnol* **26**, 317-25 (2008).

VISUAL ANALYSIS OF SPERMATOZOA, OOCYTES AND EARLY EMBRYONIC TRANSCRIPTS

Ryo Sakai[1,2*], *Ligia Mateiu*[3], *Dusan Popovic*[1,2], *Thierry Voet*[3] & *Jan Aerts*[1,2].
Dept. of Electrical Engineering (ESAT-STADIUS), KU Leuven, Belgium[1]; *iMinds Future Health Department, Belgium*[2];
Laboratory of Reproductive Genomics, Dept. of Human Genetics, KU Leuven, Leuven, Belgium [3].
[*]*ryo.sakai@esat.kuleuven.be*

Integration of heterogenous datasets from multiple experiments and different data sources to infer relationships and patterns, while considering the noise in data, is one of the important challenges accompanying modern biology. We present a design study where a visualisation tool was developed in an iterative cycle of prototype implementation and evaluation. This tool was designed to assist exploration and characterisation of zygotic and early embryonic transcripts. In close collaboration with scientific domain experts and the machine learning expert, we also applied the Self-Organizing Map clustering techniques to group transcripts with similar expression patterns. The tool aided in gaining insights during data exploration and analysis and helped development of a bioinformatics pipeline for analysis.

INTRODUCTION

Integration of heterogeneous datasets to infer important relationships and patterns while considering the noisiness of the data is one of the important challenges accompanying modern biology. In collaboration with scientific domain experts and a machine learning expert, we developed a visual analytics tool to explore and characterise the data from multi-tissue Affymetrix oligoprobe expression experiments.

METHODS

To study expression levels aiming to quantify the 3'UTR length difference of transcripts at different stages of embryogenesis, data from a type of standard 3'Affymetrix expression array across several tissues was processed at the oligo-probe level. The mapping of oligoprobes along the transcripts was done using the human genome annotation available in Biomart[1]. An initial global analysis was performed across all transcripts using the self-organizing map (SOM) algorithm[2], followed by a detailed analysis at each transcript, tissue pairwise, with analysis of covariance (ANCOVA). External datasets, such as gene ontology (GO) terms from Ensembl BioMart and Kegg pathway terms from Gene Set Enrichment Analysis[3] (GSEA), were also integrated for functional annotation.

This interactive visualization tool was developed in Processing[4], an open source programming language and integrated development based on Java. The final interface consists of multiple panels, including parallel coordinates, a scatter plot and a graph visualization of SOM clustering outputs (Figure 1). Each component was added and refined in iterative steps, based on the user feedback and hypothesises generated from prior prototypes. The interface provides focus-plus-context interactions to enable users to identify the cluster of interest and to examine individual transcripts within the selected cluster.

RESULTS & DISCUSSION

This visual analytic tool enabled the domain experts to advance the characterization of transcripts from various tissues and to develop a bioinformatics pipeline for analysis. This tool was developed in a highly flexible, rapid prototyping software development methodology to incorporate visual analysis into every step of data exploration and testing hypotheses, resulting a close feedback loop between the experts and the visualization researcher.

The overview of the expression profiles in the 3'UTRs and CDSs across tissues, generated by SOM, provides an easy access to the biologically relevant clusters. Transcripts that are tissue specific (i.e. highly expressed in one tissue while absent in the remaining tissues) or repressed in one tissue only are such examples. More complex biological scenarios (maternal transcripts decay, the activation of embryonic transcription etc.) are also identified and studied.

The future work includes developing a design methodology to guide development of low-fidelity (lo-fi) visualisation tools for biological data and a Java library to enable quick iteration and application for other domain problems.

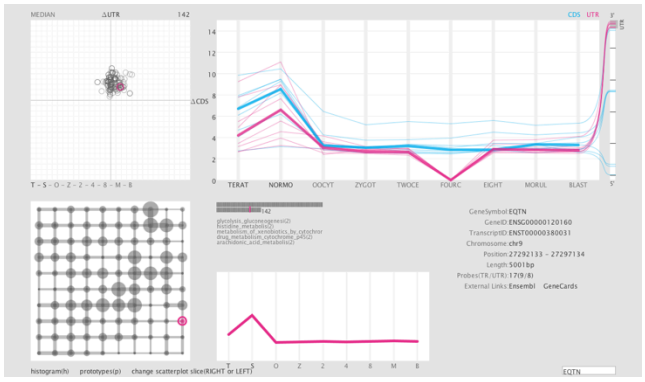


***FIGURE 1**. The interface consists of multiple panels, including scatter plot, parallel coordinates, and a graph representation of SOM clustering outputs.*

REFERENCES

1. Ensembl 2013. *Nucleic acids research* 2013, 41:D48–55.
2. Kohonen T et al. *SOM*. **6**, 1321–1344 (1997).
3. Subramanian A et al. *PNAS* **102**, 15545-15550 (2005).
4. Reas C et al. Programming Handbook for Visual Designers and Artists. *The MIT Press* (2007).

TENSOR DECOMPOSITION FOR DATA REDUCTION IN MASS SPECTROMETRY IMAGING

Yousef El Aalamat[1,2,5,*], *Raf Van de Plas*[3], *Nico Verbeeck*[1,2,5], *Etienne Waelkens*[4,5] & *Bart De Moor*[1,2,5].

Dept. of Electrical Engineering-ESAT, STADIUS Center for Dynamical Systems, Signal Processing and Data Analytics, KU Leuven[1]*; iMinds Future Health Dept, KU Leuven*[2]*; Mass Spectrometry Research Center, Vanderbilt University, Nashville, TN*[3]*; Dept. of Cellular and Molecular Medicine, KU Leuven*[4]*; Sybioma, KU Leuven*[5].

[*]*yousef.elaalamat@esat.kuleuven.be*

MALDI-based mass spectrometry imaging is rising as a prominent biomolecular imaging tool that allows study of the spatial distribution of biomolecules in a thin tissue section. This label free technique allows the detection and visualization of hundreds of molecules in a single experiment. While this property makes it a prime tool for exploratory experiments, it also results in very high-dimensional datasets and makes manual analysis of the measurements from a single tissue section difficult and impractical. In this context, data reduction techniques and statistical analysis have become indispensible for the interpretation of these data. In this work we present a dimensionality reduction technique that takes the native 3D structure of MSI data into account.

INTRODUCTION

MALDI-based mass spectrometry imaging (MSI) has been recognized as a promising biomolecular imaging tool that allows study of the abundance and the spatial distribution of known and unknown molecules in an organic tissue section. This label-free technique is a spatially resolved analytical tool permitting the detection and visualization of hundreds of molecules in a single experiment. The data resulting from an MSI experiment has a three-dimensional structure in which the xy-dimensions represent the spatial location while the third dimension represents the mass-to-charge (m/z) dimension (Figure 1). The high-dimensional nature of these datasets makes the understanding and visualization of molecular tissue compositions difficult, and usually less than exhaustive. In this context, data reduction has become an essential step in storage as well as statistical analysis.

METHODS

We implement and apply a tensor equivalent of principal component analysis (PCA) to a MSI dataset and compare the resulting patterns to the patterns extracted via standard PCA from the same dataset. The three-mode analysis yields components that are characterized by three vectors, one along each mode, rather than the two vectors provided by standard matrix PCA. We examine the applicability of this more intricate component structure towards dimensionality reduction and compressed storage of MSI data.

RESULTS & DISCUSSION

We applied the three-mode analysis methods on a MSI case study of a 10-µm thick sagittal section of the brain of a BL57/6 mouse. The MSI measurement was done using Bruker Autoflex III MALDI TOF/TOF, with the mass range extending from m/z 2800 to 25000 with 6490 m/z-bins. The grid size was 34 (x axis) by 51 (y axis) resulting in 1734 pixels with an interspot distance of 300 µm.

The three-mode analysis shows good potential for dimensionality reduction of MSI data, and shows particular applicability for compression scenarios that focus on a reduced MSI data size footprint.

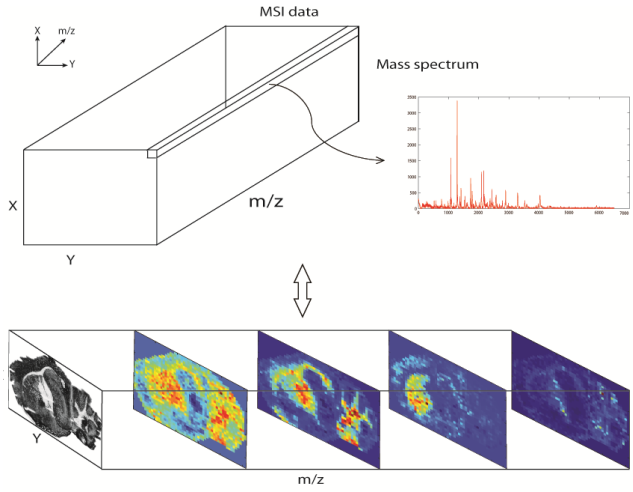


***FIGURE 1**. MSI dataset: a three-dimensional structure where the x and y dimensions describe the spatial location and the third dimension covers the chemical information axis (in m/z units).*

REFERENCES

1. De Lathauwer L. *et al. SIAM J. Mat. Anal. Appl., 21 (2000).*
2. De Lathauwer L., *"Blind Separation of Exponential Polynomials and the Decomposition of a Tensor in Rank-(Lr, Lr, 1) erms"*, *SIAM J. Mat. Anal. Appl.*, 32. 1451-1474 (2011).
3. Sorber L *et al.,"Optimization-based algorithms for tensor decompositions : canonical polyadic decomposition, decomposition in rank-$ terms and a new generalization", SIAM J. Opt., 23, 695-720(2013).*
4. Sorber L et al. Tensorlab v1.0, (2013).

LARGE-SCALE BIOIMAGE ANALYSIS USING WEB SERVICES AND MACHINE LEARNING

Raphaël Marée[1,2,*], *Benjamin Stévens*[2], *Loïc Rollus*[2], *Gilles Louppe*[2], *Louis Wehenkel*[2].
GIGA Bioinformatics Platform, Université de Liège ; GIGA Research & Dept. of Electrical Engineering and Computer Science, Université de Liège[2]. [*]*raphael.maree@ulg.ac.be*

We will present our research in bioimage analysis and our Cytomine platform [1] (http://www.cytomine.be/), a fully web-based software environment for remote visualization, collaborative annotation, and automated analysis of large-scale bioimaging datasets. It will be illustrated on various biomedical research projects involving *mouse models of inflammation-associated lung cancer*, and on toxicological and developmental studies in Zebrafish embryos.

INTRODUCTION

With recent advances in image acquisition technologies, scientists generate growing amounts of biological imaging data (e.g., in anatomical pathology, neuroscience, drug discovery, or toxicology). Projects leading to terabytes of imaging data are becoming usual in various contexts, e.g. when experimental studies rely on whole-slide virtual microscopy, high-content screening, molecular imaging by mass spectrometry, or automated volume electron microscopy. As a result, better imaging informatics tools are needed [2] to ease the visualization and high-throughput analysis of such high-dimensional datasets in today's collaborative, geographically distributed, scientific context. Indeed, as human interpretation of such datasets is impractical at such scale and operator-dependent, there is a strong need for computational methods to facilitate the extraction of quantitative information from these images.

METHODS

We have developed a rich internet application using recent web technologies and integrating various tools, standards, and machine learning and image processing algorithms. Large (> Gigabyte) bioimages can be visualized at multiple resolutions in traditional web clients through caching mechanisms and distributed image tile servers supporting various digital slide image formats. Our underlying relational data model allows to create and manage projects which contain users with permission lists, images, ontologies with domain-specific terms, and layers of annotation geometries (e.g. polygons) drawn on top of digital slide images to highlight regions of interest. All project data are stored in a spatial, relational, database and can be visualized and edited through the web interface and they can also be retrieved or updated by third-party softwares through a RESTful API. Various communication mechanisms allow multiple users to share and comment their images and annotations. In addition, image processing and recent machine learning algorithms based on randomized decision trees [3] are implemented to speed up exploration and annotation of large bioimages through content-based image retrieval, automatic object classification, and segmentation of regions of interest.

RESULTS & DISCUSSION

Our application is currently delivering about 10 000 bioimages corresponding to several terabytes of data. More than 100 000 regions of interest were annotated by ~100 scientists from multiple laboratories using ontologies describing various tissue and cell types in cancer, inflammation, and developmental studies. In particular, the proposed methodology has been recently used to identify the impact of a pulmonary tissue composition change on lung tumor onset and progression [4]. To assess these questions, different mouse models were developed where mice were treated with components inducing a specific type of neutrophilic inflammation in lung tissues. The effects of pulmonary inflammation was investigated in hundreds of lung hematoxylin-eosin-stained digital slides (each image has more than 30000 x 30000 pixels) where we used tree-based machine learning algorithms [3] to speed up the detection and quantification of tumoral regions.

The proposed framework is generally applicable and its methodological choices open the door for large-scale distributed and collaborative bioimage annotation and exploitation projects. Ongoing work includes the development of specific modules for various image-based measurements (e.g. RNAScope [5]) and their application and validation in various research projects.

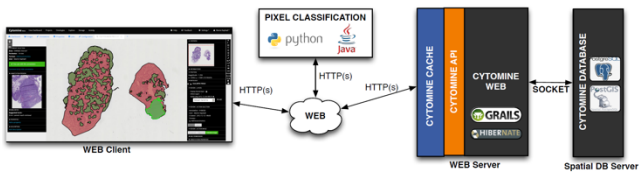


***FIGURE 1**. Overview of the Cytomine architecture with machine learning algorithms and web user interfaces to proofread algorithm predictions.*

REFERENCES

1. Marée R. et al., *BMC Diagnostic Pathology*, **8**, S26 (2013).
2. Gene Myers, *Nature Methods* **9**, 659–660 (2012)
3. Marée R. et al., invited chapter in Book Decision Forests in Computer Vision and Medical Image Analysis, Advances in Computer Vision and Pattern Recognition, Springer, 2013
4. Rocks N. et al., Poster at European Respiratory Society Annual Congress, 2013.
5. Wang F. *Journal of Molecular Diagnostics* (2012).

ACCURACY OF AN AUTOMATED VESSEL COUNTING ALGORITHM IN FOUR DIFFERENT TUMOR TYPES

Koen Marien[1,3*], Valerie Croons[3], Erik Fransen[2], Guido De Meyer[1] & Mark Kockx[3].
Dept. of Pharmaceutical Sciences[1], StatUa[2], University of Antwerp; i2, HistoGeneX NV[3]. *marien@histogenex.com

Currently, considerable attention is being paid to developing predictive biomarkers for anti-angiogenic therapies, but none exist to date. Important causes are the complex action mechanism of the therapy, study size, tumor heterogeneity, and the lack of standardized methodology. In an attempt to standardize microvessel density measurements, we compared a commercial image analysis platform (Definiens Architect) with our manual method.

INTRODUCTION

The most popular approach to measure angiogenesis is to count the smallest vessels (microvessels) in a tissue section of tumor.[1] These vessels are visible at high magnification (200x – 400x). We compared the results of our manual scoring method with the results produced by a commercial image analysis platform (Definiens Architect). Automated image analysis eliminates human subjectivity and enhances reproducibility.[2] High throughput is evident with these systems. The major limitation of image analysis relates to its accuracy, which needs to be tested and cross-validated.

METHODS

From the archives of HistoGeneX NV, 82 tissue slides were selected, all stained with the same protocol for CD31, a pan-endothelial marker used for vessel detection. A pathological report for every slide was available and used for the regional selection of tumor tissue.

The manual method consisted of random and systematic sampling using stereological techniques for the selection of a limited number of viewing fields.[3] Fifteen fields per tumor were selected. In total, 1230 fields were overlaid with a rectangular grid and analyzed for the number of vessels by two observers.

The automatic method consisted of the Blood Vessel Analysis algorithm built into Definiens Architect XD 2 (Definiens AG, Munich, Germany). The exact same area under the grid was used as for manual analysis. Four types of cancer tissue were selected: colorectal cancer (CRC), glioblastoma multiforme (GBM), ovarian cancer (OC) and renal-cell carcinoma (RCC). The settings for each tumor type were optimized, with the following CRC settings: IHC Marker = Membrane, Magnification = 20x, IHC Threshold = 0.20, Min Stain Area = 175, Gap to close = 8. Classification results (Figure 1) from the Blood Vessel Analysis algorithm were imported into Definiens Developer XD 2 (Definiens AG, Munich, Germany) to remove vessel objects that cross the left or bottom line of the grid.

In the statistical program R, a script was written to calculate the intra-class correlation coefficient (ICC) (package 'irr') and its confidence intervals (95%).[4,5] R also was used to construct plots of the prediction interval, a Bland-Altman plot and a dotplot.

RESULTS & DISCUSSION

Accuracy of the algorithm was not only dependent on the sample, but also on the tumor type (mean±SD ICC for CRC: 0.01±0.46, GBM: 0.34±0.50, OC: 0.21±0.50, RCC: -0.16±0.31). We presume that this was due to blood vessel architecture.[6] For example, when algorithm-classified objects that cross the border of the region of interest were removed, the ICC for RCC was much lower (-0.16 vs. 0.44). It is conceivable that this is because of abundant vascularization in RCC. Indeed, when border-crossing objects were removed, most vessel objects were cleared away during the image analysis. Hence, this also occurs during manual analysis, but the image analysis is stricter.

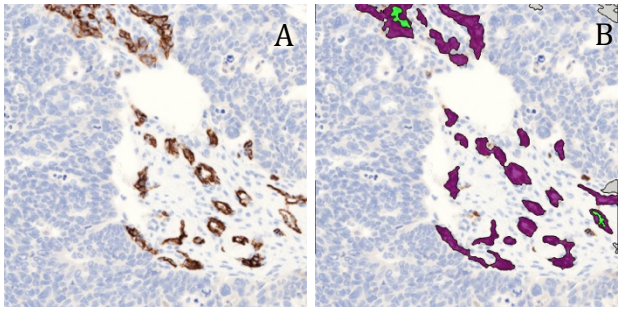


***FIGURE 1**. A) CD31-stained ovarian cancer tissue of a region of interest used for manual analysis. B) Results of the automated analysis on the same region. The algorithm detects vessels without lumen (only purple) and vessels with lumen (purple and green). Some parts of the region have been excluded for analysis (grey)*

The automatic algorithm needs to be further optimized by choosing better settings, by using different image analysis methods, and by omitting nuclear staining. It is important to perform proper artifact detection before image analysis, as false-positives and false-negatives occur throughout the series due to out-of-focus regions, folding and rupture of tissue.[7] It is also important that non-artifacts such as necrosis are identified beforehand.

REFERENCES

1. Fox, S *Methods Mol Biol* **467**, 55-78 (2009).
2. Mulrane, L *et al. Expert Rev Mol Diagn* **8**, 707-725 (2008).
3. Hansen, S *et al. Lab Invest* **78**, 1563-1573 (1998).
4. Gamer, M *et al. irr: Various Coefficients of Interrater Reliability and Agreement*. **R package version 0.84**, http://CRAN.R-project.org/package=irr (2012).
5. Kirkegaard, T *et al.Histopathology* **48**, 787-794 (2006).
6. Sabo, E *et al. Clin Cancer Res* **7**, 533-537 (2001).
7. Rizzardi, A *et al. Diagn Pathol* **7**, 42 (2012).

GUIDED EXPLORATION OF MASS SPECTROMETRY IMAGING DATA THROUGH INTEGRATION WITH ANATOMICAL INFORMATION

Nico Verbeeck[1,2,*], *Raf Van de Plas*[3], *Junhai Yang*[3], *Richard M. Caprioli*[3], *Etienne Waelkens*[4,5] & *Bart De Moor*[1,2].

Department of Electrical Engineering (ESAT), STADIUS Center for Dynamical Systems, Signal Processing, and Data Analytics, KU Leuven[1]*; iMinds Future Health Department, KU Leuven*[2]*; Mass Spectrometry Research Center, Vanderbilt University*[3]*; Sybioma, KU Leuven*[4]*; Department of Cellular and Molecular Medicine, KU Leuven*[5].

[*]*nico.verbeeck@esat.kuleuven.be*

Mass Spectrometry Imaging (MSI) is a relatively new molecular imaging technology that enables the direct analysis of the spatial distribution of molecules in a tissue section. This technology allows for the monitoring of thousands of molecules throughout the tissue in a single experiment, and enables comparison of proteomic, peptidomic, and metabolomic content between various areas in the tissue. Due to the large number of molecules that are monitored in a MSI experiment, manual investigation of these datasets typically provides only partial discovery of relevant species. Therefore, bioinformatics approaches are necessary to fully extract the relevant information from these datasets. In this work, we describe a method that enables deeper exploration by linking MSI data to an anatomical atlas.

INTRODUCTION

Recently, the significant potential of integrating MSI data with information from other data sources was demonstrated, creating richer data sets, enabling deeper insights, and combining the strengths of different imaging modalities. So far the added data sources have been limited to measured data (e.g. microscopy, MRI, etc.). In this work we demonstrate the possibilities of combining MSI data with a curated instead of a measured data source, by establishing a computational link between mass spectral measurements and an anatomical atlas. The explicit linking of ion abundance with curated anatomical labels enables a more exhaustive investigation of the roles of anatomical structures in disease and aids in data interpretation.

METHODS

Our method requires solving several sub-problems. The first task is extraction of anatomical information from the curated data source. The second task is spatial mapping of the empirically measured MSI experiment to the reference space on which the anatomical information is defined. The final task is the computational mining of the correlations between measured ion abundance patterns and curated anatomical labeling.

RESULTS & DISCUSSION

The potential of computationally linking MSI experiments with curated anatomical data sources is demonstrated through a case study on a coronal section of mouse brain. The MSI measurement was done on a Bruker AutoFlex MALDI-TOF/TOF using sinapinic acid. The acquired mass range extends from m/z 3000 to 22000, and the spatial grid consists of 53 x 98 pixels, which were measured with an interspot distance of 100 µm in both the x and y directions.

The anatomical reference space was retrieved from the publicly available Allen Brain Atlas[1] through the provided API and in-house developed code in MATLAB (The MathWorks Inc., Natick, MA). Next, a spatial mapping from physical tissue locations in the MSI experiment to locations in the anatomical reference was established through rigid and non-rigid registration techniques. Once the spatial mapping is built, the anatomical atlas annotations are explicitly linked to the MSI measurements and ready for analysis.

The computational analysis of correlations between anatomy and ion abundance can take many forms. Here, we demonstrate two types of queries: an anatomical query ("Which ions are specific for anatomical region X?"), and an ion query ("In which anatomical regions is ion Y located?"). Asking these questions yields numerous interesting correlations of ions with very specific regions in the brain, but also highlight anatomical regions where ion Y is specifically absent.

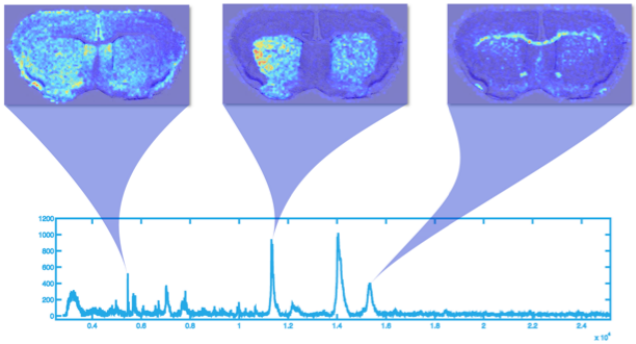


FIGURE 1. *Example of a Mass Spectrometry Dataset. Each peak in the mass spectrum represents a biomolecular ion, of which the spatial distribution in the tissue can be visualized.*

REFERENCES

1. Lein E. S. et al. *Nature* **445**, 168–176 *(*2007).

eXtasy: Variant Prioritization by Genomic Data Fusion

Alejandro Sifrim[1,2], Dusan Popovic[1,2], Leon-Charles Tranchevent[1,2], Amin Ardeshirdavani[1,2], Ryo Sakai[1,2], Peter Konings[1,2], Joris R. Vermeesch[3], Jan Aerts[1,2], Bart De Moor[1,2], Yves Moreau[1,2]

Department of Electrical Engineering, STADIUS Center for Dynamical Systems, Signal Processing and Data Analytics, KU Leuven, Leuven, Belgium[1], iMinds Future Health Department, Leuven, Belgium[2], Laboratory of Molecular Cytogenetics and Genome Research, KU Leuven, Leuven, Belgium[3]

Massive parallel sequencing greatly facilitates the discovery of novel disease genes causing Mendelian and oligogenic disorders. However, many mutations are present in any individual genome, and identifying which ones are disease causing remains a largely open problem. We introduce a new approach to prioritize nonsynonymous single nucleotide variants (nSNVs) that substantially improves prediction of disease-causing variants in exome sequencing data by integrating variant impact prediction, haploinsufficiency prediction and phenotype-specific gene prioritization.

Introduction

Next-generation sequencing (NGS) greatly facilitates the discovery of novel disease genes causing Mendelian and oligogenic disorders. However, many mutations are present in any individual genome, and identifying which ones are disease causing remains a largely open problem. We introduce a novel computational approach, called eXtasy, to prioritize nonsynonymous single nucleotide variants (nSNVs) by integrating variant impact prediction, haploinsufficiency prediction and phenotype-specific gene prioritization that allows significantly improved prediction of disease-causing variants in exome sequencing data.

Methods

To train our method we use the Human Gene Mutation Database (HGMD) as our source of disease-causing variants and 3 control sets ranging from common polymorphisms to rare variation in healthy individuals. We compare a set of commonly used classification algorithms and choose the Random Forest classifier as it outperforms all other classifiers. In order to estimate potential biases due to information leakage we perform a temporal stratification analysis.

Results & Discussion

By integrating phenotype-specific gene prioritization information we are able to greatly increase the area under the receiver-operator curve (ROC AUC) by at least 25% compared to classical deleteriousness prediction methods (e.g. SIFT, Polyphen, MutationTaster) (Figure 1). This is likely due to eXtasy's ability to discriminate between phenotype-specific and phenotype-unrelated deleterious variants. Although our performance estimates are likely overestimated due to prior information bias in a retrospective benchmark, we show that even controlling for these biases we obtain a substantial performance increase.

We believe that the presented approach will greatly facilitate the analysis of exome sequencing data in human disease by efficiently prioritizing nSNVs in the light of the phenotype in question. eXtasy is publicly available at http://homes.esat.kuleuven.be/~bioiuser/eXtasy/ (source code available at https://github.com/asifrim/eXtasy) .

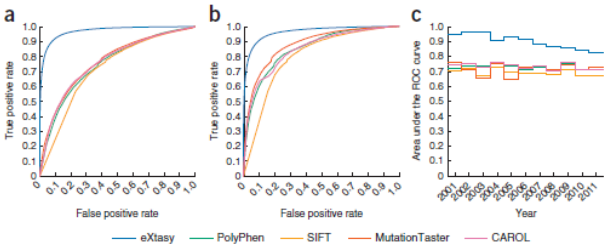


Figure 1. Receiver-Operator Curves (ROC) comparing eXtasy and classical deleteriousness prediction scores: *ROC curves for* ***(a)*** *disease-causing vs. rare control variants* ***(b)*** *disease-causing vs. common polymorphisms. In both cases eXtasy outperforms other methods as can be seen by an increase in the area under the curve (AUC).* ***(c)*** *To test the effect of biases in our retrospective benchmark we compared obtained AUCs by stratifying disease-causing variants on the year of discovery. More recent disease variant associations show a decrease in performance for eXtasy as biases decrease. The method however always outperforms classical deleteriousness prediction scores.*

References

Sifrim, A. *et al. Nature methods* (2013). doi:10.1038/nmeth.2656

PREDICTION ACCURACY FOR DELETERIOUS AND DISEASE CAUSING MUTATIONS IN HEALTHY INDIVIDUALS

Raf Winand[1,2,*], *Kristien Hens*[3], *Alejandro Sifrim*[1,2], *Inge Liebaers*[4], *Joris Vermeesch*[5], *Yves Moreau*[1,2] & *Jan Aerts*[1,2].

Department of Electrical Engineering (ESAT), STADIUS Center for Dynamical Systems, Signal Processing and Data Analytics, KU Leuven[1] *; iMinds Future Health Department*[2] *; Health, Ethics & Society, Maastricht University*[3] *; Centre for Medical Genetics, University Hospital Brussels*[4] *; Laboratory of Cytogenetics and Genome Research, Centre for Human Genetics (CME), KU Leuven*[5]. *raf.winand@esat.kuleuven.be*

Whole genome sequencing comes with the unprecedented opportunity to obtain secondary information not related to the original clinical question. In order for this information to prove useful, assessing the clinical validity is crucial. We must ensure that prediction algorithms and databases are able to accurately predict the phenotype in healthy individuals. To this end we compared the specificity of several prediction algorithms and databases like SIFT, PolyPhen and HGMD regarding the detection of damaging or disease causing mutations. We also tested a new in-house developed algorithm called eXtasy that is disease-centric and incorporates the phenotype of the individual in the analysis.

INTRODUCTION

The last decade has seen a tremendous evolution in the field of human genetics. With the decreasing cost of next-generation sequencing methods it is likely that a physician will order a full genome test instead of a single gene test. This whole genome sequence however contains secondary information that is not necessarily related to the original clinical question. Many algorithms exist that predict whether a specific mutation will have a damaging effect on the protein level. Also curated databases that hold disease associated and disease causing mutations exist. But are these algorithms and databases able to accurately predict the phenotype in healthy individuals?

METHODS

We selected a list of 347 congenital disorders characterized by extreme dysmorphologies, early onset, and a known molecular basis from OMIM[1]. We analyzed samples from people who are considered healthy from the 1000 Genomes Project[2], publicly available samples from Complete Genomics[3], and in-house samples from our university. In those samples, we looked for mutations in genes listed in OMIM as associated with the selected disorders. For all mutations predictions were made with both SIFT[4] and PolyPhen[5]. We also looked for mutations that are annotated as disease causing in HGMD[6]. In addition we tested our new in-house developed algorithm called eXtasy[7] that incorporates the phenotype in the analysis and is disease-centric.

RESULTS & DISCUSSION

A mutation was considered severe when it was either predicted to be damaging by both PolyPhen and SIFT, an essential splice site mutation or a nonsense mutation. Our analysis shows that 40-94% and 2-69% of individuals, although considered 'healthy', have at least one severe mutation with a minor allele frequency of <1% (in the thousand genomes) in genes associated with respectively autosomal dominant and autosomal recessive diseases.

When looking at mutations that are annotated as disease causing in HMGD, we found that 12-22% and 0-8% of samples showed mutations associated with respectively autosomal dominant and autosomal recessive diseases. These mutations however are associated with a limited number of disorders suggesting a limited penetrance or spurious annotation.

To compare the performance of eXtasy with the other prediction algorithms we calculated the score for the most occurring nonsynonymous mutations that are predicted to be damaging by both PolyPhen and SIFT and compared those values against the eXtasy scores of known disease causing mutations. eXtasy shows fewer false positives as can be seen in Figure 1. The scores of the analyzed mutations here are still rather high albeit on the low side of the HGMD mutations.

In all datasets we found numerous mutations that are predicted to be damaging and/or disease causing. Many healthy individuals carry these mutations in genes associated with severe disease. Although these predictably damaging mutations show low eXtasy scores compared to the HGMD mutations, there is no clear separation between the two. In order to be useful for the analysis of secondary genomic information, improvements still have to be made. Because of the many splice-site variants we encountered, prediction scores for these variants would be of great value.

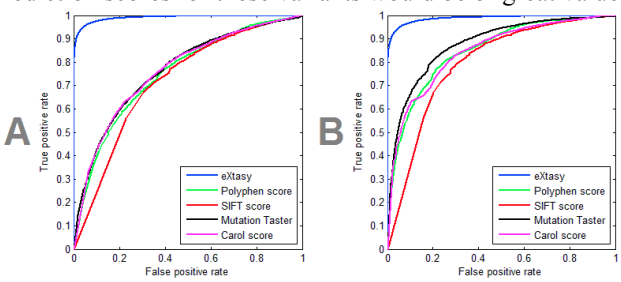


***FIGURE 1**. Receiver-Operator Curves comparing eXtasy and other popular prediction scores for (A) disease causing vs. rare control variants and (B) disease causing vs. common polymorphisms.*

REFERENCES

1. McKusick-Nathans Institute of Genetic Medicine JHU. Retrieved from http://omim.org Accessed 2013-04-02.
2. McVean G a *et al. Nature* **491**, 56–65 (2012).
3. Drmanac R *et al. Science* **327**, 78-81 (2010).
4. Ng PC *Nucleic Acids Res* **31**, 3812-3814 (2003)
5. Adzhubei I a *et al. Nat Methods* **7**, 248-249 (2010)
6. Stenson PD *et al. Hum Mutat* **21**, 577-581 (2003)
7. Sifrim A *et al. Nat. Methods* ***Epub ahead of print*** (2013)

CONVERT YOUR FAVORITE PROTEIN MODELING PROGRAM INTO A MUTATION PREDICTOR: "MODICT"

Ibrahim Tanyalçin[1,2*], Danny Coomans[3], Katrien Stouffs[1,4], Willy Lissens[1,4], Anna C. Jansen[2,5]

*Centre for Medical Genetics, UZ Brussel, Brussels, Belgium[1], Neurogenetics Research Group, Vrije Universiteit Brussel, Brussels, Belgium[2], Department of Biomedical Statistics and Informatics, Vrije Universiteit Brussels, Belgium[3], Reproduction, Genetics and Regenerative Medicine, Vrije Universiteit Brussel, Belgium[4], Pediatric Neurology Unit, Department of Pediatrics, UZ Brussel, Brussels, Belgium[5]. *itanyalc@vub.ac.be*

As next generation sequencing is dominating the field of research today with the accelerated data processing through bioinformatics, more and more missense SNPs and stopgain, stoploss or frame-shift mutations are being identified than ever before. As of today, a variety of mapping and filtering tools are already available for rapidly converting raw data into output, but unfortunately the rate of generating raw data greatly exceeds the rate of its analyzing. One of the greatest challenges is to be able to predict whether these variations are real disease causing changes behind one's condition.

INTRODUCTION

The current concept of mutation prediction heavily depends on the integrated algorithms that mainly implement a sequence based BLAST search that tries to identify a number of similar protein sequences above a preset threshold, than relate and combine several other parameters such as PSIC (**P**osition-**S**pecific **I**ndependent **C**ounts), known 3D structures of similar proteins, surface area, β-factor and atomic contacts. **However there are no algorithms that can quantitatively extract information from server predicted 3D protein models and provide a relative assessment for the probability of the mutation being damaging**.

METHODS

Algorithm

Initially, the information from .pdb (**P**rotein **D**ata **B**ank), and server predicted 3D models have to be extracted in means of RMSD (**R**oot **M**ean **S**quare **D**eviation) and compared. For this purpose we have derived an algorithm consisting of 3 important steps, initial raw score, a significance measure and a final score:

$$\textbf{Final Score} = \frac{\left(1+\frac{\sum Diff-\sum Sig}{|\sum Diff|+|\sum Sig|}\right)}{2} \times \frac{\sum A.D}{2\sqrt{\left(\frac{\sum A.D}{\sum Total}\right)^2+\left(\frac{\sum A.E}{\sum Total}\right)^2}}$$

Generation of 3D protein models – for mutations in FOXG1, RENIN and TUBB2B:

All models used were generated by submitting raw *foxg1*, *renin* and *tubb2b* wildtype and mutated sequences to the automated I-Tasser server. Resulting models were energy minimized via 2 cycles of steepest descent consisting of 50 steps each and 1 cycle of conjugate gradient consisting of 200 steps with a minimum ΔE of 0.01kJ/mol together with a harmonic constraint of 100 kJ/mol. Models were further refined using the ModRefiner (Xu and Zhang 2011). For each query a trio pair was constructed by comparison of the ratio of the final scores between wildtype/wildtype^{refined}, wildtype/test and wildtype/mutated where the foremost and the lattermost components serve as negative and positive controls respectively. All renderings and modifications were performed on DeepView–Swiss-PdbViewer (http://www.expasy.org/spdbv/). Images were post-processed with POV-Ray v3.6.

RESULTS & DISCUSSION

All of the features of the chosen mutations were correctly predicted by the algorithm. Mutations that were less accurately predicted in the databases were initially chosen. Below is the demonstration of 3D protein models and algorithm score comparison of 2 mutations on the same protein, ren^{R33W} (benign) and ren^{C20R} (deleterious) respectively:

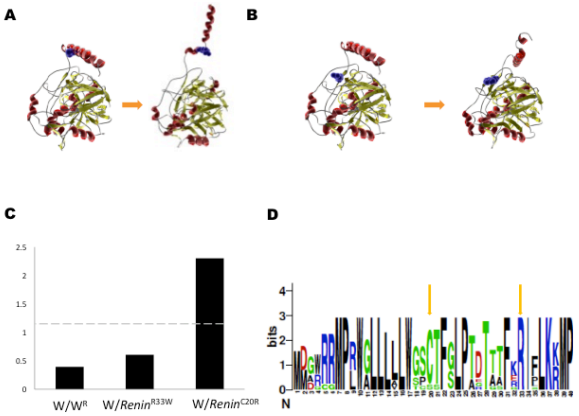


***FIGURE 1*. 3D models of wildtype and mutated *Renin*. A.** Wildtype (right) and *ren* $^{p.C20R}$ (left) with cysteine and arginine residues rendered with blue Van der Waals radii. Alpha helixes are colored in red whereas beta strands are colored in yellow. **B.** A benign polymorphism in the signal sequence, *ren*$^{p.R33W}$ (right) does not result in a change to the same extent as *ren* $^{p.C20R}$. **Graphical representation of algorithm scores (C)**. Absolute values of algorithm scores obtained from pairs; negative control (left; score: ~0.395), wildtype against *ren*$^{p.R33W}$ (middle; score: ~0.609) and positive control (right, wildtype against *ren*$^{p.C20R}$. Score: ~2.304). Dashed lines (light gray) mark half the difference between the scores of positive and negative controls. **Sequence logo of the *renin* signal peptide (D).** Residues from 1-40 of reviewed *renin* sequences in UniProt database have been aligned. Note that both R33 and C20 are highly conserved, however algorithm scores significantly differ.

In the near future we are planning to fully automate the score generation process, making it easier and faster to utilize for scientific community.

REFERENCES

1. Roy A. et al. *Nat Protoc,* **5**, 725-38 (2010).

UNRAVELLING THE REGULATORY MECHANISMS BEHIND INTER-GENIC CARDIAC QUANTITATIVE TRAIT LOCI THROUGH SYSTEMS GENETICS APPROACHES

Michiel E. Adriaens[1], Tamara T. Koopmann[1], Perry D. Moerland[2], Margriet L. Westerveld[1], Roos F. Marsman[1], Sean Lal[3], Taifang Zhang[4], Christine Simmons[5], Istvan Baczko[6], Christobal dos Remedios[3], Nanette H. Bishopric[4,7], Al L. George jr.[5], Andras Varro[6], Connie R. Bezzina[1]*

*[1]Department of Experimental Cardiology, Heart Failure Research Centre, Academic Medical Center, Amsterdam, The Netherlands; [2]Department of Clinical Epidemiology, Biostatistics and Bioinformatics, Academic Medical Center, Amsterdam, The Netherlands; [3]Muscle Research Unit, Department of Anatomy, Bosch Institute, The University of Sydney, Sydney, Australia; [4]Department of Medicine, University of Miami School of Medicine, Miami, FL, USA; [5]Departments of Medicine and Pharmacology, Vanderbilt University, Nashville, TN, USA; [6]Department of Pharmacology and Pharmacotherapy, Faculty of Medicine, University of Szeged, Hungary; [7]Department of Molecular and Cellular Pharmacology, University of Miami School of Medicine, Miami, FL, USA. *m.e.adriaens@amc.uva.nl*

A considerable portion of chromosomal loci associated with cardiac traits (GWAS) lie in inter-genic regions[1]. Trait-associated SNPs located in these putative regulatory regions (e.g. promoters en enhancers) likely exert their effect by modulating gene expression[2]. The key to unravelling molecular mechanisms underlying cardiac traits is to interrogate these variants for association with differential transcript abundance by expression quantitative trait locus (eQTL) analysis.

INTRODUCTION

In recent years, multiple genome-wide eQTL resources for various tissues such as for brain, liver and adipose tissue have been made available. Because eQTLs may be tissue specific a similar resource for human heart is anticipated to have great value[3-6]. We therefore set out to perform the first genome-wide eQTL analysis of the non-diseased human heart.

METHODS

A total of 129 left ventricle samples were collected from non-diseased donor hearts. Gene expression of over 47000 transcripts and genotypes of 1.3 million SNPs were determined using Illumina microarray technology. After pre-processing and stringent quality control, each transcript was tested for association with all SNP genotypes using linear modelling in R (GenABEL).

RESULTS & DISCUSSION

We identified 793 independent eQTLs mapping to 389 unique transcripts (FDR<5%). Over 99% of these are cis interactions (eQTL SNP located within 1 Mb of a gene). An example is the effect of rs9912468, a robust modifier of ventricular depolarization time, on PRKCA, providing evidence for the role of this protein as the effect mediator (**Figure 1**). We are currently exploring the regulatory mechanisms behind eQTL effects by overlaying with cardiac regulatory regions identified through ChIP-sequencing, i.e. binding regions of cardiac specific transcription factors (TBX3/TBX5/NKX2-5/GATA4), enhancers (p300) and epigenetic markers of open chromatin (H3K4me3/H3K27ac/H3K9ac).

This study is an illustration of integrating gene expression, phenotype and genotype datasets through systems genetics approaches, to hunt down novel causal genes for human cardiac phenotypes.

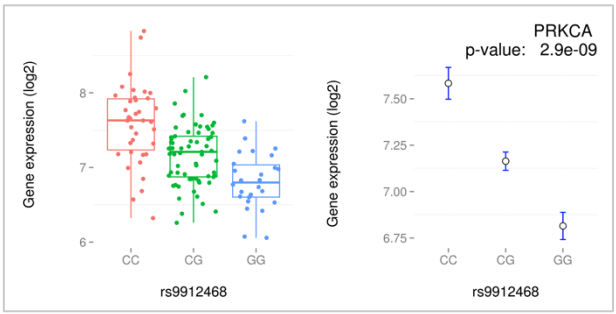


***FIGURE 1**. An overview of a GWAS cis eQTL: rs9912468 with PRKCA. On the left of the panel, box-and-whisker plots of mRNA levels for all genotypes. On the right, mean and standard-error plots of mRNA levels for all genotypes are illustrated.*

REFERENCES

1. Cookson, W., Liang, L., Abecasis, G., Moffatt, M. & Lathrop, M. Mapping complex disease traits with global gene expression. *Nat Rev Genet* **10**, 184-194, doi:10.1038/nrg2537 (2009).
2. Nicolae, D. L. *et al.* Trait-associated SNPs are more likely to be eQTLs: annotation to enhance discovery from GWAS. *PLoS genetics* **6**, e1000888, doi:10.1371/journal.pgen.1000888 (2010).
3. Grundberg, E. *et al.* Mapping cis- and trans-regulatory effects across multiple tissues in twins. *Nature genetics* **44**, 1084-1089, doi:22941192 (2012).
4. Hernandez, D. G. *et al.* Integration of GWAS SNPs and tissue specific expression profiling reveal discrete eQTLs for human traits in blood and brain. *Neurobiology of disease* **47**, 20-28, doi:10.1016/j.nbd.2012.03.020 (2012).
5. Nica, A. C. *et al.* The architecture of gene regulatory variation across multiple human tissues: the MuTHER study. *PLoS genetics* **7**, e1002003, doi:21304890 (2011).
6. Petretto, E. *et al.* Heritability and tissue specificity of expression quantitative trait loci. *PLoS genetics* **2**, e172, doi:10.1371/journal.pgen.0020172 (2006).

CONNECTING PHENOTYPES AND TRAITS TO BIOLOGICAL PROCESSES AND MOLECULAR FUNCTIONS

Aalt D.J. van Dijk[1,2,*], *Jan-Peter Nap*[1], *Gabino F. Sanchez-Perez*[1].
Bioscience[1], *and Biometris*[2], *Wageningen University and Research Centre.* [*]*aaltjan.vandijk@wur.nl*

We provide a large scale inventory of connections between phenotypes and gene functions (biological processes and molecular functions), based on integration of QTL and GWAS data with gene function predictions. Various approaches, including text mining, are applied for validation of our predictions. The predicted connections provide insight into the underlying molecular biology of traits and phenotypes. For example, for various human diseases, as well as for plant traits related to yield, gene functions that are involved are predicted. In addition, we demonstrate that we can prioritize candidate genes underlying traits.

INTRODUCTION

A major limitation of existing studies of the genetic basis of phenotypes/traits is that such studies do not directly provide access to the molecular and mechanistic basis of traits. Although Quantitative Trait Loci (QTLs) or Genome Wide Association Studies (GWAS) indicate genetic loci underlying a trait, they do not directly uncover the underlying biological processes or molecular functions. Here we provide a first large scale inventory of links between traits and gene functions, obtained based on overrepresented function in QTL/GWAS regions.

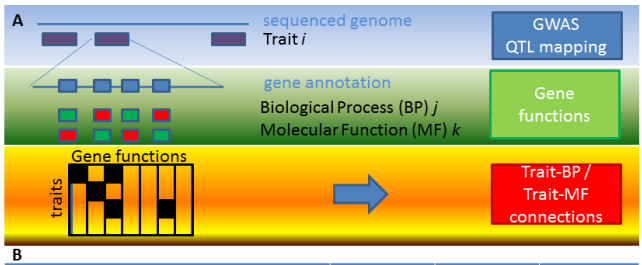
METHODS

We assessed overrepresentation of predicted gene functions (both biological processes and molecular functions as defined in the Gene Ontology) in regions associated to traits according to several QTL/GWAS datasets (Figure 1A): *(i)* human disease GWAS compendium; *(ii)* Arabidopsis GWAS dataset; *(iii)* rice QTL compendium; *(iv)* Arabidopsis metabolomics-QTL (mQTL) dataset. For the biological processes, we compare the use of purely sequence-based gene function prediction with our integrative sequence- and network-based gene function prediction[1,2]. The resulting links between traits and biological processes were validated in various ways, including simulations and text mining. The performance of our procedure to pinpoint relevant candidate genes was assessed by comparison with candidate gene fine-mapping results.

RESULTS & DISCUSSION

A clear difference was observed between the different types of traits in terms of preferential linking to molecular functions or biological processes (Figure 1B). The metabolite traits were connected almost exclusively with molecular functions, and not with biological processes. For the other plant trait datasets, this was largely the reverse (mostly biological processes were associated to these traits); whereas for the human disease dataset, molecular function and biological process links were both found at large numbers. Importantly, for the biological processes linked to the human and plant traits, we demonstrate a clear added value of using a network-based method for function prediction that we previously developed[1], compared to alternative purely sequence-based strategies of gene function annotation: the number of statistically significant links between biological processes and traits was much higher with our predicted functions (not shown).

For the metabolite traits, the majority of the predicted molecular functions could easily be recognized as relevant. For example, these functions included enzymatic functions related to the metabolites, as well as transporters. Text mining indicated that for the non-metabolite traits, the biological processes that we predicted to be linked, occur much more often together in Pubmed abstracts than would be randomly expected. This leads to validation of many of our trait-biological processes predictions. For the human dataset, this includes for example links between cholesterol related diseases and cholesterol related gene functions, and between metabolic syndrome and metabolic processes. For the rice dataset, this includes non-trivial predictions such as a link between the trait root number and terpene biosynthetic processes, between leaf senescence and starch related processes, and between potassium uptake and glycogen/glucan related processes. Intriguingly, for several Arabidopsis flowering related traits, links to defense related biological processes are predicted. Finally, comparison with QTL fine-mapping results indicates a promising performance of our approach to prioritize putative candidate genes underlying traits of interest.



B

Dataset	#traits	#BP links	#MF links
Human GWAS compendium	661	3842	9890
Arabidopsis GWAS dataset	107	218	24
Rice QTL compendium	147	1080	342
Arabidopsis mQTL dataset	156	6	1487

FIGURE 1.(A) *Integrating gene function predictions with QTL or GWAS data leads to links between phenotypes and biological processes or molecular functions.* ***(B)*** *Number of traits and number of predicted links per dataset.*

REFERENCES

1. Kourmpetis et al., Bayesian Markov Random Field analysis for protein function prediction based on network data. PloS ONE 5 (2), e9293, 2010.
2. Bargsten et al, *submitted to PLoS Comp Biol.*

A CLASS REPRESENTATIVE MODEL FOR PURE PARSIMONY HAPLOTYPING UNDER UNCERTAIN DATA

Luciano Porretta[1*], *Martine Labbé*[1], *Daniele Catanzaro*[2]
[1]*Department of Computer Sciences, Université Libre de Bruxelles (U.L.B.), CP 210/01, Brussels, Belgium;* [2]*Department of Operations, Rijksuniversiteit Groningen, 9700 AV, Groningen, the Netherlands.* [*] *lporrett@ulb.ac.be*

The Pure Parsimony Haplotyping (PPH) problem is a NP-hard combinatorial optimization problem that consists of finding the minimum number of haplotypes necessary to explain a given set of genotypes. PPH has attracted more and more attention in recent years due to its importance in analysis of many fine-scale genetic data. Its application fields range from mapping complex disease genes to inferring population histories, passing through designing drugs, functional genomics and pharmacogenetics. In this article we investigate, for the first time, a recent version of PPH called *the Pure Parsimony Haplotype problem under Uncertain Data* (PPH-UD). This version mainly arises when the input genotypes are not accurate, i.e., when some single nucleotide polymorphisms are missing or affected by errors. We propose an exact approach to solution of PPH-UD based on an extended version of Catanzaro *et al.* [1] class representative model for PPH, currently the state-of-the-art integer programming model for PPH. The model is efficient, accurate, compact, polynomial-sized, easy to implement, solvable with any solver for mixed integer programming, and usable in all those cases for which the parsimony criterion is well suited for haplotype estimation.

INTRODUCTION

The human genome is divided in 23 pairs of chromosomes thereof, one copy is inherited from the father and the other from the mother. When a nucleotide site of a specific chromosome region shows a variability within a population of individuals then it is called *Single Nucleotide Polymorphism* (SNP). A *haplotype* is a set of SNPs and represent a fundamental source of information for disease association studies. In fact, over 90% of sequence variation among individuals is due to common variant sites, most of which arose from single historical mutation events on the ancestral chromosome [2]. Hence, in a group of people affected by a disease, the SNPs causing or associated with the disease will be enriched in frequency compared with the corresponding frequencies in a group of unaffected individuals. Extracting haplotypes from a population of individuals is not an easy task. In fact, the current molecular sequencing techniques only provide information about the conflation of the paternal and maternal haplotypes of an individual (also called *genotype*) rather than haplotypes themselves [3]. When the family-based genetic information of a population is available, haplotypes can be retrieved experimentally [4]. However, the experimental approach is generally laborious, cost-prohibitive, requires advanced molecular isolation strategies [5], and sometimes not even possible [6]. In absence of a family-based genetic information, a valid alternative to the experimental approach is provided by computational methods which estimate, by means of specific criteria, haplotypes from the set of genotypes extracted from a population of individuals.

RESULTS & DISCUSSION

In this article we proposed an exact approach to solution of PPH-UD based on an extended version of Catanzaro *et al.* [1] class representative model for PPH, possibly one of the best integer programming models described so far in the literature on PPH. The model is efficient, accurate, compact, polynomial-sized, easy to implement, solvable with any solver for mixed integer programming, and usable in all those cases for which the parsimony criterion is well suited for haplotype estimation.

REFERENCES

1. Catanzaro D, et al. *INFORMS Journal on Computing* **22**, 195–209 (2009).
2. Li WH & Sadler LA. *Genetics* **129**, 513–523 (1991).
3. Halldórsson BV. et al. In: Calude CS, editor. Discrete Mathematics and Theoretical Computer Science, Springer-Verlag, volume 2731 of *Lecture Note in Computer Science* (2003).
4. Lu X. et al. *Genome Res*. **13**, 2112–2117 (2003).
5. Clark VJ. et al. *Human Genetics* **108**: 484–493 (2001).
6. Lancia G. et al. *INFORMS Journal on Computing* **16**: 348–359 (2004).

CLASSIFYING THE PROGRESSION OF DUCTAL CARCINOMA FROM SINGLE-CELL SAMPLED DATA : A CASE STUDY

Daniele Catanzaro[1], Salim A. Chowdhury[2], Stanley E. Shackney[3], and Russell Schwartz[2].*
*[1]Dep. of Operations, Rijksuniversiteit Groningen, 9700 AV, Groningen, the Netherlands; [2]Dep. of Human Oncology and Human Genetics, University of Pittsburgh, School of Medicine, PA2; [3]Dep. of Biological Sciences, Carnegie Mellon University, 4400 Fifth Avenue, Pittsburgh, PA3. *Correspondence should be addressed to: d.catanzaro@rug.nl*

Ductal carcinoma in situ is a precursor lesion of invasive ductal carcinoma of the breast. Investigating its temporal progression could provide fundamental new insights for the development of more effective diagnoses and treatments. In this work we address this major issue by investigating the problem of reconstructing a plausible progression of the carcinoma from single-cell sampled data of an affected individual. Specifically, by using a number of assumptions derived from the observation of cellular atypia occurring in ductal carcinoma, we design a possible predictive model based on the parsimony criterion. Preliminary experiments carried out on a population of 13 patients show that the corresponding predicted progressions are non-random and classifiable in subfamilies having specific evolutionary characteristics.

INTRODUCTION

Ductal Carcinoma In Situ (DCIS) is considered a precursor lesion for invasive breast cancer and is found synchronously in approximately 45% of patients affected by *Invasive Ductal Carcinoma* (IDC)[1]. Specifically, DCIS is the last step in a continuum of non- invasive stages of increased cellular atypia, which are believed to develop from flat epithelial atypia and atypical ductal hyperplasia[2]. The incidence of both DCIS and IDC is currently estimated in 35 and 155 per 100,000 women in the United States, respectively, with an increasing outlook possibly caused by the improvement of the accuracy of diagnoses[3]. Investigating the temporal progression of this type of carcinoma could provide fundamental new insights for the development of more effective diagnoses and treatments, hence increasing research efforts have been devoted to this task in recent years[4]. A possible attempt aiming to understand the dynamics of genomic alterations during the evolution of the carcinoma has been performed recently by Heselmeyer-Haddad et al.[5]. Specifically, the authors carried out a single-cell analysis[6,7] on 13 patients affected by invasive ductal carcinoma of the breast and observed both an enormous intercellular heterogeneity in DCIS and IDC (although lower in DCIS with respect to IDC) and signal patterns consistent with a non-random distribution of genomic imbalances. The presence of recurrent patterns of genomic imbalances in the evolution from DCIS to IDC led the authors to suspect the existence of a precise sequence of genetic events causing the progression from DCIS to IDC. The authors however did not investigate the topic any further.

Starting from Heselmeyer-Haddad et al.[5] results, in this work we address the problem of modeling and classifying the progression of ductal carcinoma in a population of affected individuals. Specifically, by making use of a number of assumptions derived from the observation of cellular atypia occurring in ductal carcinoma, we first design a possible parsimony- based predictive model able to reconstruct a plausible progression of the carcinoma from single-cell samples of an affected individual. Subsequently, we test the predictive model on the same population of 13 patients described in Heselmeyer-Haddad et al.[5]. Our preliminary experiments show that the corresponding predicted progressions are non- random and classifiable in subfamilies having specific evolutionary characteristics

REFERENCES

1. von Minckwitz G. et al. *Breast Cancer Research and Treatment*, **132**, 863–870 (2012).
2. Sgroi DC. *Annual Reviews of Pathology* **5**, 193–221 (2010).
3. Virnig BA et al. T. *Journal of the National Cancer Institute* **102**, 170–178 (2010).
4. Podlaha O. et al. *Cell* **28**, 155–163 (2012).
5. Heselmeyer-Haddad K. et al. *American Journal of Pathology*, **181**, 1807–1822 (2012).
6. Pennington G. et al. *Journal of Bioinformatics and Computational Biology* **5**, 407–427 (2006).
7. Pennisi E. *Science* **336**, 976–977 (2012).

COMPREHENSIVE ANALYSIS OF TRANSCRIPTOME VARIATION UNCOVERS KNOWN AND NOVEL DRIVER EVENTS IN T-CELL ACUTE LYMPHOBLASTIC LEUKEMIA

*Zeynep Kalender Atak[#*1], Valentina Gianfelici[#2,3], Gert Hulselmans[#1], Kim De Keersmaecker[#2], Arun George Devasia[1,2], Ellen Geerdens[2], Nicole Mentens[2], Sabina Chiaretti[3], Kaat Durinck[4], Anne Uyttebroeck[5], Peter Vandenberghe[2], Iwona Wlodarska[2], Jacqueline Cloos[6], Robin Foà[3], Frank Speleman[4], Jan Cools[2], and Stein Aerts[1].*

*Laboratory of Computational Biology, Center for Human Genetics, KU Leuven, Leuven, Belgium[1];Laboratory for the Molecular Biology of Leukemia Center for Human Genetics, KU Leuven and Center for the Biology of Disease, VIB, Leuven, Belgium[2];Division of Hematology, Department of Cellular Biotechnologies and Hematology, 'Sapienza' University of Rome, Italy[3];Center for Medical Genetics, Ghent University, Ghent, Belgium[4];Pediatric Hemato-Oncology, University Hospitals Leuven, Leuven, Belgium[5]; Pediatric Oncology/Hematology and Hematology, VU Medical Center, Amsterdam, The Netherlands[6];equal contribution[#]. * zeynep.kalender@med.kuleuven.be*

RNA-seq data provides a rich data source that can be used to interrogate different modalities of genomic alterations inherent in cancer samples. We have used RNA-seq on T-cell acute leukemia (T-ALL) genomes to identify novel driver genomic events. T-ALL is an aggressive hematological malignancy caused by gene fusions leading over-expression of transcription factors and mutations causing aberrant cellular signaling. By using RNA-seq we uncovered not only known driver genomic aberrations but also novel candidate driver events such as fusions *SSBP2-FER* and *JAK2-TPM3*; or mutations in *H3F3A*, *STAT5*, *PTK2B* and *CTCF*.

INTRODUCTION

We have used RNA-seq on 31 diagnostic samples and 18 cell lines with the aim of obtaining accurate gene expression levels, identifying single nucleotide variants (SNV) and small insertions and deletions (INDEL), detecting alternative transcript events (ATE) as well as gene fusions.

METHODS

First, to identify accurate gene expression levels we have implemented rigorous normalization and batch effect removal procedures. Next, we validated and optimized variant calling pipelines for RNA-seq by using our previously published exome dataset[1]. We have identified high quality mutation predictions by implementing extensive filtering and obtained candidate driver genes by recurrence based analysis and gene prioritization techniques[2]. Next, to detect alternative transcript events we performed *de novo* transcript assembly and interrogated the assembled transcripts for existence of novel transcript structures in the known T-ALL driver genes. Finally, we have identified fusion events by making use of the paired-end structure of the RNA-seq data, and obtained high quality predictions by implementing additional filters.

RESULTS & DISCUSSION

Distinct molecular subtypes of T-ALL are associated with aberrant expression of certain marker genes. We have demonstrated that RNA-seq data can be successfully used for this task as aberrant expression of these marker genes classified samples into correct subtypes in our dataset. Moreover, we have used expression data throughout the analysis to evaluate findings identified in the proceeding analysis steps.

Mutation detection in RNA-seq proved to be heavily dependant on the mapping strategy, and by using matched RNA-seq data and Exome-seq data, we determined the best mapping strategy would be a combined mapping strategy consisting of mapping reads to genome, to transcriptome and then to genome again as split reads[3]. This mapping strategy together with subsequent variant calling, filtering and prioritization steps; we obtained mutations in many established driver genes such as *NOTCH1*, *FBXW7*, and *BCL11B*; but also in promising novel candidates such as *H3F3A*, *STAT5*, *PTK2B* and *CTCF*.

Alternative transcript discovery analysis resulted in detection of two novel exon skipping events in known T-ALL drivers *SUZ12* and *LCK*. Overall relatively few ATEs were identified thus T-ALL provided a robust transcript isoforms usage.

Fusion discovery resulted in identification of known events in T-ALL such as *STIL-TAL1* and *NUP214-ABL1* fusions, but also novel events resulting in chimeric oncoproteins such as *SSBP2-FER* and *TPM3-JAK2* fusions, or in aberrant expression of a known T-ALL driver such as *TLX1*, *PLAG1*, *LMO1* or *NKX2-1*.

In conclusion, we present comprehensive analysis of the T-ALL transcriptome by evaluating the gene expression perturbations, mutations, alternative transcript events and gene fusions.

REFERENCES

1. de Keersmaecker, K. *et al.* Exome sequencing identifies mutation in CNOT3 and ribosomal genes RPL5 and RPL10 in T-cell acute lymphoblastic leukemia. *Nat Genet* **45,** 186–190 (2013).
2. Aerts, S. *et al.* Gene prioritization through genomic data fusion. *Nature Biotechnology* **24,** 537–544 (2006).
3. Kim, D. *et al.* TopHat2: accurate alignment of transcriptomes in the presence of insertions, deletions and gene fusions. *Genome Biol* **14,** R36 (2013).

DETECTING MASTER REGULATORS AND CIS-REGULATORY INTERACTIONS IN HUMAN CANCER RELATED GENE NETWORKS

Rekin's Janky[1]*, *Annelien Verfaillie*[1], *Gert Hulselmans*[1] *and Stein Aerts*[1]
Laboratory of Computational Biology, Department of Human Genetics, KU Leuven, Belgium[1].
rekins.janky@med.kuleuven.be

Gene regulatory networks play key roles in cancer, and cancer driver mutations are over-represented in transcription factors and associated transcriptional regulators. Cancer genome sequencing has delivered many new candidate drivers, but their downstream effects are often unknown. Therefore identification of direct interactions between transcription factors (TFs) and their target genes is of central importance to gain insight into oncogenic programs. Here we describe a powerful computational method, called iRegulon, for the identification of "regulons" in a cancer gene signature. iRegulon is available as a Cytoscape plugin (http://iregulon.aertslab.org), for the first time combining advanced cis-regulatory sequence analysis with network biology.

INTRODUCTION

The identification of master regulators of a biological process, and understanding their effect on gene expression in the cell, is a very broadly applicable challenge. A regulon consists of a TF and its direct transcriptional targets, which contain common TF binding sites in their cis-regulatory control elements. We present iRegulon, the first motif discovery based method to discover, analyze, and annotate human and mouse regulons in a gene list or in an existing gene network or signaling pathway.

METHODS

The iRegulon plugin allows you to identify regulons using motif discovery in a set of co-regulated genes. The prediction of regulons consists of the following three steps.

(A) Motif detection, consisting of two parts. The first part is the offline scoring of a sequence search space (up to 20kb around the TSS) around every gene in the human genome using a Hidden Markov Model, detecting homotypic motif clusters (cluster-buster[1]). This is done for more than six thousands position-weight matrices (PWM), resulting in a gene-ranking for each PWM. This process is repeated for the orthologous sequences in ten other vertebrate genomes, followed by the integration of the cross-species rankings using rank aggregation. The second part of motif detection is the on-the-fly identification of those motifs for which the input genes are enriched at the top of the ranking, using the Area Under the Curve (AUC) of the cumulative recovery curve. Enriched motifs are those with a high AUC compared to the average AUC of all motifs, and enrichment is measured by a normalized enrichment score (NES).

(B) Motif2TF mapping: the prioritization of candidate TFs that could bind to the enriched motifs. This is achieved by finding the optimal path from a motif to a TF, in a motif-TF network where the edges consist of motif2motif similarity, TF2TF orthology, and motif2TF annotation.

(C) Target detection: The determination of the optimal subset of direct target genes, namely the significantly highly ranked genes compared to the genomic background and to the entire motif collection as background.

The final output is thus a list of enriched motifs, together with a prioritized list of candidate transcription factors that can bind the motif, and for each motif a set of direct target genes. The cytoscape plugin works as a java client connected to the server-side daemon over the internet. The iRegulon server-side daemon is implemented in Python and uses MySQL to store and query the motif-based whole-genome rankings. After submitting a gene set or network to the service, the results are returned to the client (this happens on-the-fly, and takes about one minute). The user can browse through the motif discovery results, select a TF among the prioritized list of TFs, and add upstream regulators and direct regulator-target 'edges' to the input gene set or network under study.

RESULTS & DISCUSSION

We extensively validate iRegulon on many different data sets, including the entire ENCODE ChIP-Seq compendium, and we clearly describe its sensitivity and specificity with respect to regulator and target discovery. Next, we find that iRegulon markedly outperforms eight other motif discovery tools. We illustrate how gene regulatory networks can be inferred from TF perturbation signatures, microRNA target sets, signaling pathways and protein-protein interaction networks. Next, we perform RNA-seq and ChIP-seq experiments and map a functional p53-dependent network in breast cancer cells, identifying new p53 co-factors, target genes and binding sites, which we validate experimentally. We show how iRegulon can be a valid alternative to ChIP-Seq, and how it can be used jointly with ChIP-Seq and RNA-Seq. Finally, we apply iRegulon to twenty thousand cancer gene signatures and propose a new concept of a meta-regulon, representing a transcription factor and its 'context-free' targetome.

REFERENCES

1. Frith, M. C. *et al. Nucleic acids research* **31,** 3666–8 (2003).

LONG NON-CODING RNAS IN LUNG CANCER: COMPARISON OF MICROARRAY AND RNA-SEQ TECHNIQUES

Petr V. Nazarov[1,*], *Tony Kaoma*[1], *Arnaud Muller*[1], *Sabrina Fritah*[2] & *Laurent Vallar*[1].
Genomics Research Unit[1] *and Neuro-Onclolgy Laboratory*[2], *Centre de Recherche Public de la Santé*
[*]*petr.nazarov@crp-sante.lu*

Here we address the expression of long non-coding RNA (lncRNA) in non-small cell lung cancer (NSCLC) and characterize differences in the results based on two experimental approaches: exon-level microarrays and next-generation RNA sequencing (RNA-seq). We performed several types of statistical and bioinformatics analyses for lncRNA expression in NSCLC patient samples, including differential expression, co-expression and genomic co-location.

INTRODUCTION

Only a 2% of the human genome is protein-coding, while 75% of it can be transcribed[1]. Recently this part of transcriptome has dragged an attention as it was shown that non-coding sequences have an important regulatory role. While the role of short non-coding RNAs, such as microRNA, is already well characterized, the area of lncRNA was actively investigated only since recent years.

Here we compare two experimental approaches – standard microarray technology and next-generation RNA-seq – for the analysis of changes in lncRNA expression profiles in NSCLC cancer tissue samples.

METHODS

Tumour and matched normal tissues were collected in the framework of Luxembourg FNR CORE grant "ASTSTRO" from 8 patients with adenocarcinoma (AC) and 10 patients with squamous cell carcinoma (SCC) and processed by standardized pipelines.

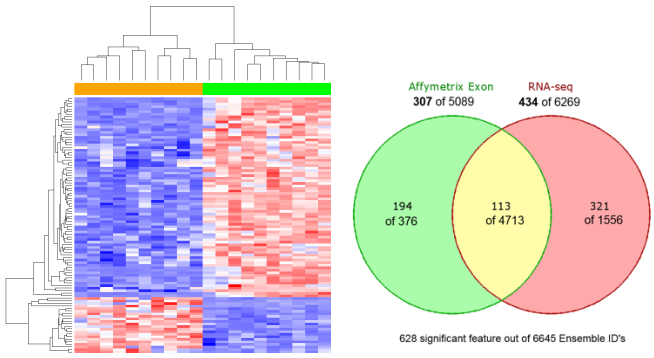
We used Affymetrix® Human Exon 1.0 ST arrays and Illumina® paired-end RNA-seq sequencing to profile RNA expression in matched normal adjacent tissue and tumor pairs. Microarray data were preprocessed using GC-RMA and differentially expressed lncRNAs were detected by *limma* package of R/Bioconductor. RNA-seq data were processed using the TopHat-Cufflinks suite, while significance analysis was performed by processing sequencing data with EdgeR.

For annotation of lncRNAs we used *biomaRT* package of R/Bioconductor.

Co-expression analysis of lncRNA and mRNAs was performed using CoExpress[2] tool (www.bioinformatics.lu) taking into account all available samples simultaneously.

RESULTS & DISCUSSION

Although the average expression of lncRNA was approximately twice lower than of protein-coding mRNA, we were able to identify 307 differentially expressed lncRNAs for Affymetrix arrays with FDR<0.01 (top 100 are shown in Figure1A) in SCC. RNA-seq analysis returned 434 annotated lncRNA for SCC. We found only ~30% of significant lncRNAs common in both analyses, while overlap based on annotated features was more than 80%. Due to high variability of normal samples paired with AC, no differentially expressed lncRNAs were found for AC group by microarray technique and only 26 were detected by RNA-seq (FDR<0.01).



A B

FIGURE 1*. (A) Expression of top 100 significant lncRNA changing between SCC cancer (orange) and normal lung tissue (green) based on microarray data. (B) Overlap between differentially expressed and annotated lncRNA detected from microarray and RNA-seq analysis.*

Correlation analysis of interactions inside lncRNA data showed high bias to positive correlations. This bias might be an artefact linked to microarray normalization strategy and suggests that lncRNA data extracted from microarray experiments should be rectified by additional between-array normalization. After between-array normalization analysis of lncRNA:mRNA co-expression returned mostly positive correlations. These positive co-expressions can be explained by co-location of mRNA and lncRNA in a genome.

Interestingly, we found in microarray data features corresponding to a well-known non-coding prognostic marker for lung cancer MALAT1[3], which showed constantly high expression. At the same time no reads were found for this transcript in RNA-seq data. Closely located SCYL1 mRNA was found constantly expressed by both techniques.

As expected, our study showed that RNA-seq offers benefits compared to microarrays for the analysis of non-coding RNAs. It holds especially to detection and characterization of new lncRNAs. However, we observed some evidences (e.g. detection of MALAT1 transcript) that microarray may outperform RNA-seq in some situations for detection of well-known lncRNAs.

REFERENCES

1. Djebali S, *et al. Nature* **489**, 101-108 (2012).
2. Nazarov P, *et al. Nuc. Acid Res.* **41**, 2817-2831 (2013).
3. Ji S, *et al. Oncogene* **22**, 8031–8041 (2003).

EXTRACTING SIGNATURES FROM HIGH-DIMENSIONAL UNBALANCED BIOLOGICAL DATA: THE CASES OF DNA METHYLATION AND LNCRNA IN BREAST CANCER

Martin Bizet[1-2,*], *Matthieu Defrance*[2], *Olivier Van Grembergen*[2], *Sarah Dedeurwaerder*[2], *François Fuks*[2] & *Gianluca Bontempi*[1]

Machine Learning Group[1] *and Laboratory of Cancer Epigenetics*[2], *Université Libre de Bruxelles.* [*]*mbizet@ulb.ac.be*

In clinical and biological fields it is often essential to select a subset of variables able to accurately discriminate between two conditions (*e.g.* healthy *versus* cancerous). Since the advent of the high-throughput technologies (micro-arrays and next-generation sequencing) the number of variables available grows exponentially and discovering this subset – called a signature – becomes challenging. Because dealing with this high-dimensionality and with the unbalancement of the data is particularly critical, we developped a machine-learning based R-pipeline able to extract signatures while taking these two considerations into account. This pipeline has been successfully applied to investigate two oncological questions: could we predict response to a chemotherapy using DNA methylation data and could the long non-coding RNAs help to discriminate cancers from healthy samples.

INTRODUCTION

While clinicians need to make accurate prediction of some critical parameters (*e.g.* patient response to a chemotherapy), biologists want to discover key-candidates that help to understand specific questions (*e.g.* the mechanism of a disease). The use of machine learning approaches is a good solution in both cases. The principle is to extract the subset of variables – called a signature - that best discriminate between two groups (or more). However some difficulties remains to reach a good prediction power.

Firstly, the "curse of dimensionality" - the impossibility to make accurate predictions when the dimensionality (i.e. the number of variables) is too high - is a well-known problem in the machine learning field[1] and has become an important issue with the introduction of high-throughput technologies. Secondly, biological data are often very unbalanced (for practical reasons) with one group being under-represented compared to the other. This could also heavily impact the prediction accuracy.

METHODS

We developed a R-pipeline to extract a signature from high-dimensional unbalanced data (Figure 1):

Input: high-dimensional unbalanced data

Processing:

- First an unsupervised feature selection step is applied to roughly reduce the dimensionality.
- Then a three-fold cross-validation is repeated 200 times. It includes 4 substeps:
  - A second feature selection that limit the dataset to a small dimensionality
  - A balancement of the data
  - The training and test of a supervised classifier algorithm
  - The prediction accuracy computation
- As a negative control the same step is also applied on randomized data.
- Prediction accuracy from randomized and real data are compared.
- The final signature is computed.

Ouptut: a set of predictive variables.

RESULTS & DISCUSSION

Breast cancer is the woman first cause of cancer death in occidental countries[2], so predicting the response to a chemotherapy is of first interest. To investigate the role of DNA methylation for such a question we used 51 non-responders and 7 responders samples (collaboration with Bordet Institute). The samples were profiled using Illumina HumanMethylation450 (more then 450 000 variables). After preprocessing[3], our pipeline was applied. We obtained a good prediction power (Matthew Correlation Coefficient, MCC > 0.35) which is significantly beter than random (t-test p-value = 5.7e-47). The signature is currently under investigation.

This R-pipeline was also applied on long non-coding RNA data to discriminate normals and breast cancer tissues: 9 normals and 45 cancer samples were investigated using a custom micro-array interrogating the expression of more the 23 000 non-coding RNAs. Again a good prediction power was obtained using our R-pipeline (MCC > 0.85; t-test p-value < 1e-100). Experimental evaluation of candidates is ongoing.

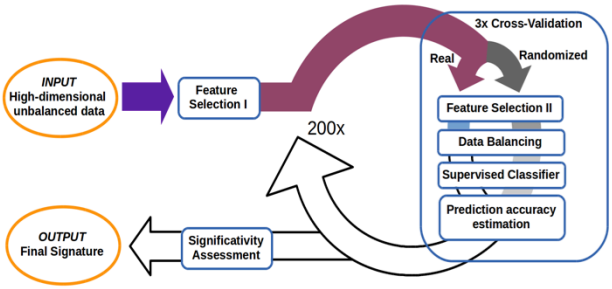


***FIGURE 1**. Workflow to extract a signature from high-dimensional unbalanced data (available as a R-pipeline).*

REFERENCES

1. Bellman R. Adaptative Control Processes: A Guided Tour. Princeton University Press, Princeton, New Jersey 1961.
2. Dedeurwaerder S. et al *EMBO Mol. Med.* **3**,726-41 (2011)
3. Dedeurwaerder S. et al *Brief. Bioinformatics* **Epub** (2013).

BELLEROPHON: A HYBRID METHOD FOR DETECTING INTERCHROMOSOMAL REARRANGEMENTS AT BASE PAIR RESOLUTION USING NEXT-GENERATION SEQUENCING DATA

Matthew Hayes[1,2,*], *Jing Li*[1,*]
Department of Electrical Engineering and Computer Science[1], *Case Western Reserve University*
Department of Electrical Engineering and Computer Science[2], *McNeese State University*
[*]*mhayes2@mcneese.edu, jingli@case.edu*

Somatically-acquired translocations may serve as important markers for assessing the cause and nature of diseases like cancer. Algorithms to locate translocations may use next-generation sequencing (NGS) platform data. To address the challenge of finding and classifying translocation breakpoints, we have developed "Bellerophon", a method that uses discordant read pairs to identify potential translocations, and subsequently uses "soft-clipped" reads to predict the location of the precise breakpoints. Using two simulated datasets and two prostate cancer datasets, Bellerophon had favorable performance. Our method also accurately predicted the presence of the interchromosomal insertions placed in our simulated dataset, which is an ability that the other SV prediction programs lack. Because it does not perform assembly on soft-clipped subreads, Bellerophon may be limited in experiments where sequence read lengths are short.

INTRODUCTION

Somatic genomic structural variants (SV) are highly associated with onset and susceptibility to diseases such as cancer[1]. These variants may inactivate genes that are critical to the prevention of cancer onset, or they may amplify "oncogenes", which leads to increased susceptibility. Oncogenic gene fusions can also occur due to the presence of SVs[2]. Interchromosomal structural variants include reciprocal and non-reciprocal translocations, and also interchromosomal insertions, in which one chromosome donates a contiguous segment to a non-homologous chromosome[3]. These variants can also be highly associated with the presence of certain cancer types. Given the potential association of some translocations with cancer, it is important to develop computational methods to locate and classify them at base pair level. We developed "Bellerophon" to address these challenges.

METHODS

Like most programs for SV detection, Bellerophon takes as input a set of read alignments in BAM format and looks for dense clusters of abnormally mapped or *discordant* read pairs. For interchromosomal variants, the discordant pairs of interest are the ones that are *chimeric*; one read maps to some chromosome *i*, and its mate maps to a non-homologous chromosome *j*. After finding these clusters of chimeric read pairs, the program looks for *soft-clipped* reads that are within the cluster. If the clipped subread remaps to opposite cluster, then the cluster is predicted as an interchromosomal variant. After predicting precise breakpoints, the method proceeds to the classification stage. Bellerophon will attempt to classify each breakpoint as a participant in an **unbalanced translocation, balanced translocation,** or **interchromosomal insertion**.

RESULTS & DISCUSSION

In our experiments, we created two simulated genomes containing various interchromosomal SVs. We simulated the sequencing of these genomes and aligned the sequence reads to the human reference hg18. We also aligned the sequence reads of two primary prostate cancer cell lines and aligned these with BWA. We compared the performance of our method to four SV detection programs. For both simulated datasets, Bellerophon correctly identified and classified all breakpoints. Bellerophon detected structural variants with greater precision than the paired end methods, and greater sensitivity on the low-coverage datasets than a popular split-read method for SV detection. Furthermore, Bellerophon will attempt to classify interchromosomal variants as 1) balanced translocations, 2) unbalanced translocations, or 3) interchromosomal insertions.

In conclusion, we have presented Bellerophon, a method for identifying and classifying interchromosomal rearrangements at base-pair resolution. Bellerophon had similar breakpoint prediction ability to CREST[4], and it called fewer false positives than the paired-end methods on the cancer datasets. Future work will address limitations such as 1) dependence on soft-clipping and 2) focusing only on interchromosomal.

REFERENCES

1. R. Redon et al. "Global variation in copy number in the human genome." *Nature* 2006, **444:**444-454.
2. F. Mitelman et al. "The impact of translocations and gene fusions in cancer causation.". *Nature Reviews Cancer* 2007, **7:**233-245.
3. Gardner R, Sutherland G. *Chromosome Abnormalities and Genetic Counseling*. Oxford, UK: Oxford University Press; 1989.
4. J. Wang et al. "CREST maps somatic structural variation in cancer genomes with base-pair resolution." *Nat Methods* 2011, **8:**652-654.

A HUMAN-SPECIFIC ENDOGENOUS RETROVIRAL INSERT SERVES AS AN ENHANCER FOR THE SCHIZOPHRENIA-LINKED GENE PRODH

Maria Suntsova[1,2], Elena Gogvadze[1], Sergey Salozhin[3], Nurshat Gaifullin[4,5], Fedor Eroshkin[1], Sergey E. Dmitriev[6], Natalia Martynova[1], Kirill Kulikov[5], Galina Malakhova[1], Gulnur Tukhbatova[3], Alexey P. Bolshakov[3], Dmitry Ghilarov[1], Andrew Garazha[1,2], Alexander Aliper[1,2], Charles Cantor[7], Yuri Solokhin[5], Pavel Balaban[3], Alex Zhavoronkov[2], Anton Buzdin[1,2].*

Group for Genomic Regulation of Cell Signaling Systems[1], Shemyakin-Ovchinnikov Institute of Bioorganic Chemistry, Moscow, Russia; Laboratory of Bioinformatics[2], D. Rogachyov Federal Research Center of Pediatric Hematology, Oncology and Immunology; Laboratory of Molecular Neurobiology[3], Institute of Higher Nervous Activity and Neurophysiology; Faculty of Fundamental Medicine[4], Lomonosov Moscow State University; The Russian National Research Medical University named after N.I. Pirogov[5]; Belozersky Institute of Physico-Chemical Biology[6], Lomonosov Moscow State University; Department of Biomedical Engineering[7], Boston University, Boston, Massachusetts, United States of America. *suntsova86@mail.ru

Understanding the molecular basis of phenotypic differences between humans and chimpanzees can provide important clues to human specific behavioral peculiarities and neurological disorders. Using a systematic, whole-genome analysis of enhancer activity of human-specific endogenous retroviral inserts (hsERVs), we identified a previously-unknown element, hsERV$_{PRODH}$, that acts as a tissue-specific enhancer for the PRODH gene, required for proper central nervous system (CNS) functioning.

INTRODUCTION

Retroelements occupy ~45% of the human genome[1] and contain various regulatory sequences, such as promoters, enhancers and polyadenylation signals and are thus strong candidates for a role of functional genome reshapers. HsERVs of the HERV-K(HML-2) group are one of the four families of transposable elements that were active at the time of human and chimpanzee ancestor divergence[2].

METHODS

In this study we performed full genome bioinformatic survey of 133 human-specific copies of HERV-K(HML-2). Six of them were mapped in the close neighborhood (closer than 5 kb) to transcription start sites of known human genes. Enhancer activity of the corresponding insertions was studied using reporter assay. Epigenetic regulation of hsERV$_{PRODH}$ and CpG$_{PRODH}$ enhancer activity was performed using bisulfite sequencing, methyl sensitive high resolution melting assay and *in vitro* methylation. *PRODH* transcription was examined using quantitative PCR and public available gene expression data. Search of transcription factors involved in hsERV$_{PRODH}$ regulation was conducted with microarray data analysis followed by bioinformatic mapping of putative transcription factor binding sites. Interaction of transcription factor SOX2 with respective binding sites within hsERV$_{PRODH}$ was analyzed by electrophoresis mobility shift assay. We also performed computational analysis of SOX2 recognition motive representation among the HERV-K(HML-2) family members.

RESULTS & DISCUSSION

PRODH is one of the candidate genes for succeptibility to sczophrenia and other neurological disorders[3]. It codes for a proline dehydrogenase enzyme, which catalyses the first step of proline catabolism and is most likely involved in neuromediator synthesis in the CNS. Here, we show that *PRODH* is positively regulated by a human-specific endogenous retroviral (hsERV) insert. The hsERV$_{PRODH}$ enhancer strongly up-regulates *PRODH* promoter activity in a tissue-specific manner. HsERV $_{PRODH}$ enhancer activity is regulated by methylation, and is highest when in the unmethylated state. In contrast, for another *PRODH* regulatory region – CpG$_{PRODH}$ - methylation status seems to be completely independent of transcriptional status. We hypothesize that the hsERV$_{PRODH}$ insertion at some point in human evolution may have significantly influenced *PRODH* transcriptional activity, and increased its expression in the CNS. We show that the hsERV$_{PRODH}$, together with the *PRODH* promoter, drives neuron-specific expression in cultured hippocampal cells. Remarkably, hsERV$_{PRODH}$ was hypomethylated in human hippocampal samples, where *PRODH* showed the highest transcriptional activity. We also uncovered a mechanism for the regulation of hsERV $_{PRODH}$, which involves the binding of SOX2. The role of SOX2 in regulating the *PRODH* gene and also endogenous retroviruses in general was not previously known. We propose that the interaction of hsERV$_{PRODH}$ and *PRODH* may have contributed to human CNS evolution.

REFERENCES

1. Lander ES. et al. Initial sequencing and analysis of the human genome. *Nature* 409**:** 860-921 (2001).
2. Mills RE. et al. Recently mobilized transposons in the human and chimpanzee genomes. *Am J Hum Genet* 78**:** 671-9 (2006).
3. Kempf L. et al. Functional polymorphisms in PRODH are associated with risk and protection for schizophrenia and fronto-striatal structure and function. *PLoS Genet* 4**:** e1000252 (2008).

IN SILICO DRUG REPURPOSING IN PARKINSON'S DISEASE

Patrice Godard[1] , Matthew Page[2] & Jonathan van Eyll[2, *].
Thomson-Reuters IP & Sciences[1]; UCB[2]. *jonathan.vaneyll@ucb.com

Costly, protracted drug development, coupled with increasingly high clinical attrition rates, have fuelled the pharmaceutical industry's interest in drug repurposing strategies. Drug repurposing is the rational application of a known drug to new indications and can lead to shorter, less costly drug development cycles with increased probability of success. Parkinson's Disease (PD) is the 2nd most common neurological disorder, involving progressive disruption of motor function with possible psychiatric complications, caused by depletion of dopaminergic neurons in the nigrostriatal system. Drug repurposing represents a potential strategy for cost saving and risk mitigation in the treatment of neurodegenerative diseases, including PD, where drug approval rates are low (~8%). A data-driven strategy was adopted to identify drugs in clinical development that target molecules relevant to the patho-physiology of PD but that are not indicated for PD. Systems-level genomics data sets were combined with commercial knowledge management resources to enable a rational prioritisation of repurposing candidates.

INTRODUCTION

It is estimated to take 10 to 17 years to develop a drug de novo[1], with an associated cost of over $900 million[2]. Costly protracted drug development coupled with increasingly high attrition rates for drugs in clinical development have raised the interest for drug repurposing strategies in pharmaceutical industry. Drug repurposing (or repositioning) is the application of a known drug to new indications (e.g. new diseases) other than those for which it was originally intended. Apart from the opportunity to create value, drug repositioning has a number of research and development advantages including shortening development timelines by up 3 to 5 years, elimination or deep reduction of pre-clinical research costs and, in many cases, lowering attrition rates due to availability of clinical safety data. Re-purposed compounds are known to have higher success rates in trials[3].

Along with oncology (5%), central nervous system (CNS) disease is one of the therapeutic areas with the lowest success rates for drug approval of 8%[2]. Thus, the economic risk to develop therapies for CNS diseases is significantly higher than for other diseases. Repositioning existing drugs for these indications could reduce drug development risk and increase productivity in this specific therapeutic area. In this frame, *in silico* approaches show a strong potential benefit.

Parkinson's disease (PD) is the second most common neurodegenerative disease (the first one is Alzheimer disease). In Western countries, PD affects 1 to 2% of the population older than 65 and the number of cases is expected to double by 2030[4,5]. To date there is no cure for Parkinson's disease, although several treatments are available to manage the symptoms[5].

The aim of this study was to identify targets and their corresponding drugs which do not belong to the current treatment landscape of Parkinson's disease but which could show a benefit in this frame.

METHODS

Three kinds of approaches have been chosen in order to achieve this goal. The first one is based on available knowledge of the molecular mechanisms underlying the disease, such as biomarkers or pathways. The second one is based on the identification of key actors in the transcriptional regulatory network underlying the disorder. The third, "guilty by association" approach relies on pathway and biomarker similarity between PD and other diseases.

All these approaches have been integrated in different classifiers. Besides the identification of putative candidate drugs for Parkinson's disease repurposing, the performance of the different prediction methods and classifiers has been assessed according to specific controls.

RESULTS & DISCUSSION

The final classifier for prediction of repurposing targets identified 126 candidates for PD. Among them, 67 are already studied in the context of this disease, 32 being the positive controls used to train the model. At the time being, the 59 remaining targets are not under the focus of any project related to Parkinson's disease. 26 of them are targeted by marketed drugs or close to launch, 10 are studied in the frame of early of discontinued projects, and 23 are related to project, at most, under clinical studies of phase III (Figure 3). The later targets are of particular interest because of the available data related to their safety and tolerability without being commercialized yet. After reviewing them, the most promising will be tested in vivo.

REFERENCES

1. Tobinick, E.L. (2009). The value of drug repositioning in the current pharmaceutical market. Drug News Perspect. 22, 119–125.
2. Kola, I. & Landis, J. (2004). Can the pharmaceutical industry reduce attrition rates? Nat Rev Drug Discov 3, 711–715.
3. Ashburn, T.T. & Thor, K.B. (2004). Drug repositioning: identifying and developing new uses for existing drugs. Nat Rev Drug Discov 3, 673–683.
4. Thomas, B. (2011). Molecular insights into Parkinson's disease. F1000 Medicine Reports 3.
5. Stacy, M., Hickey, P. & Stacy, M. (2011). Available and emerging treatments for Parkinson's disease: a review. Drug Design, Development and Therapy, 241.
6. Krauthammer, M., Kaufmann, C.A., Gilliam, T.C. & Rzhetsky, A. (2004). Molecular triangulation: bridging linkage and molecular-network information for identifying candidate genes in Alzheimer's disease. Proc. Natl. Acad. Sci. U.S.A. 101, 15148–15153.

ASSESSMENT OF REFERENCE NETWORKS FOR PATHWAY ANALYSIS AND MECHANISTIC INTERPRETATION OF DISEASE DATA

Ana Carolina Fierro[1,2,*], *Bo Colruyt*[3,4], *Sofie Van Landeghem*[3,4], *Yves Van de Peer*[3,4], *Kathleen Marchal*[1,2,4].

Center of Microbial and Plant Genetics, Kasteelpark Arenberg 20, B-3001, Leuven, Belgium [1]; *Department of Information Technology, Ghent University, IMinds, 9052 Gent, Belgium*[2]; *Department of Plant Systems Biology, VIB, 9052 Ghent, Belgium* [3]; *Department of Plant Biotechnology and Bioinformatics, Ghent University, 9052 Ghent, Belgium* [4].

[*]*Carolina.Fierro@biw.kuleuven.be*

Network-based analyses are currently used to gain insights into the mechanisms underlying diseases. Several databases provide partial views on diverse interaction types but the choice of data sources is rather arbitrary in each case study. For mechanistic interpretation purposes, a trade-off exists on the reference network: curated interactions provide high reliability information but they are largely incomplete, while experimental and predicted interactions are able to link more genes among them but with a much lower reliability. In this study we evaluate several sources of interactions and their performance on sub-network methods. We provide guidelines to select a reference network that best suits network-based methods to perform mechanistic interpretation of human disease data.

INTRODUCTION

Network-based analyses combined with high-throughput data are currently used to gain insights into the mechanisms underlying diseases. For instance, network-based methods have been used to identify deregulated pathways in diseases, driver genes and pathways in cancer and to predict novel disease-associated genes. These methods require as input a list of molecular interactions that build the network. Several databases provide partial views on diverse interaction types, but the human interactome is far from being fully characterized and many predicted interactions are available. Currently most network-based studies rely on curated networks and they mainly compile ad-hoc reference networks from public databases. Thus, the choice of a reference network is rather arbitrary. However, it has been shown that the human network chosen as reference does influence network-based methods performance[1]. Existing benchmark studies have focused on disease gene prioritization methods, which aim at predicting new genes associated to a particular disease regardless of the molecular mechanism behind it. In contrast, the identification of deregulated sub-networks or path-finding methods aim to elucidate the molecular mechanism responsible for deregulated gene expression in diseases or finding the pathways that link genotype to molecular phenotypic response.

For mechanistic interpretation purposes, a trade-off exists on the reference network: curated interactions provide high reliability information but they are largely incomplete. In such case, lack of data might leave undetected many important genes and interactions involved in a particular process or disease. On the other hand, experimental and predicted interactions are able to link more genes among them but with a much lower reliability, and wrongly predicted edges render the mechanistic interpretation more difficult.

METHODS

In this study we evaluate several sources of interactions, ranging from highly curated to functional human networks. First we provide an overview of the network properties of each network and then we evaluate the connectivity they provide with respect to known disease pathways. Secondly we evaluate the performance of sub-network identification[2,3] on small scale datasets. Since most network-based methods rely on gene expression data as input, we use breast data from The Cancer Genome Atlas to perform our analysis on a real dataset.

RESULTS & DISCUSSION

We provide guidelines to select a reference network that best suits network-based methods to perform mechanistic interpretation of human disease data.

REFERENCES

1. Gonçalves JP *et al*. *PLoS One* **7(11)**, e49634 (2012).
2. Ulitsky I *et al*. *Plos One* **5(10)**, e13367 (2010).
3. Beisser D *et al*. *Bioinformatics* **26(8)**,1129-30 (2010).

INTEGRATED ANALYSIS OF TRANSCRIPT LEVEL REGULATION OF METABOLISM REVEALS DISEASE RELEVANT NODES OF THE HUMAN METABOLIC NETWORK

Mafalda Galhardo[1#], Lasse Sinkkonen[1#], Philipp Berninger[2], Jake Lin[3,4], Thomas Sauter[1] and Merja Heinäniemi[1,5*].*

*[1]Life Sciences Research Unit, University of Luxembourg, 162a Avenue de la Faïencerie, L-1511 Luxembourg, Luxembourg. [2]Biozentrum, Universität Basel and Swiss Institute of Bioinformatics, Klingelbergstrasse 50-70, 4056 Basel, Switzerland. [3]Institute for Systems Biology, 401 Terry Avenue North, 98109-5234, Seattle, Washington, USA. [4]Luxembourg Centre for Systems Biomedicine, University of Luxembourg, L-4362 Esch/Alzette, Luxembourg. [5]A. I. Virtanen Institute for Molecular Sciences, University of Eastern Finland, FI-70120 Kuopio, Finland. # = equal contribution, *Thomas.Sauter@uni.lu, merja.heinaniemi@uef.fi*

Metabolic diseases and co-morbidities represent an ever-growing epidemic where multiple cell types impact tissue homeostasis. Here, the link between the metabolic and gene regulatory networks was studied through integrated experimental and computational analysis of public ChIP-seq data from HUVEC and gene expression and regulatory data from SGBS adipocyte cells. On both datasets, disease association was found on metabolic genes with higher regulatory load. An open-source web portal, IDARE (Integrated Data Nodes of REgulation), was established for visualizing the various gene-related data on HUVEC and SGBS cells, in context of metabolic pathways.

INTRODUCTION

The aim of this study was to depict disease-relevant links between the human gene regulatory and metabolic networks through integrative analysis of two cellular systems widely used to study complex metabolic diseases.

METHODS

HUVEC dataset: Public ChIP-seq data from the genome-wide binding profile of 10 transcription factors (TFs) on HUVEC cells were used to overlay the regulatory and metabolic networks. The list of target genes from all TFs was used to fill gene metanodes that were coupled to Recon1[1] metabolic pathways. Gene-disease associations were extracted from the DisGeNet[2] database and association enrichment was assessed via hypergeometric tests on genes bound by different numbers of the 10 TFs.

SGBS dataset: SGBS pre-adipocytes were converted into lipid-loaded adipocytes following a 12-day differentiation process. Using microarrays, we obtained the time-course expression of genes and microRNAs (miRNAs). ChIP-seq for PPARγ, CEBPα, LXR and the H3K4me3 were performed to identify genome-wide putative targets of the three TFs and genes with the H3K4me3 active TSS mark. MiRNAs miR-27a, miR-29a and miR-222, consistently down-regulated during adipogenesis, were selected for the identification of their putative targets via transcriptomic profiling and seed enrichment analysis following specific overexpression of each miRNA. The expression of metabolic genes was used as input for the constraint-based method of Shlomi et al.[3], to predict the activity of metabolic reactions in Recon1. The diverse data were overlaid on Recon1 metabolic pathways, with gene-related data depicted through gene metanodes and reaction activity prediction difference between pre-adipocytes and adipocytes shown by edge colors. For the interactive visualization of the data here collected, the web portal **IDARE** (http://systemsbiology.uni.lu/idare.html) was established, allowing to easily perceive links between the gene regulatory and metabolic networks.

RESULTS & DISCUSSION

In the HUVEC dataset, we observed a trend of high regulatory load on genes relevant for disease, namely NOS3 and MTAP, both associated to vascular diseases, GOT2 associated with cardiac arrhythmia and ALDH4A1 with type II hyperprolinemia, all four putatively bound by at least 8 TFs. Hypergeometric enrichment tests for endothelial relevant disease association were statistically significant for genes with at least 6 TFs (Figure 1), suggesting that highly regulated genes are more prone to be relevant for disease.

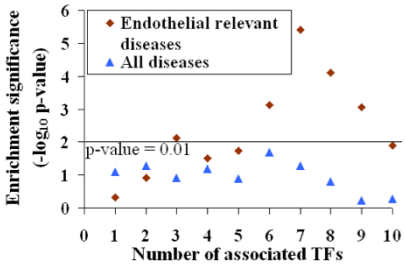


***FIGURE 1**. Gene-disease association enrichment p-value for different TF numbers.*

In the SGBS dataset, we observed the convergence of regulatory load from multiple TFs and miRNAs on lipid metabolism related genes with increased expression, suggesting a combinatorial role for the regulators on driving metabolic changes characteristic of adipocyte differentiation. Known lipodystrophy genes were also readily exposed by our integrative approach. The BCAA degradation was the non-lipid pathway predicted to change most in adipocytes with miRNA regulation on initial key steps.

Our work exemplifies the benefit of taking an integrative approach for the analysis of multiple omics data to depict links between biological networks with disease relevance.

REFERENCES

1. Duarte et al., *PNAS*, 2007, doi: 10.1073/pnas.0610772104.
2. Bauer-Mehren et al., *Bioinformatics*, 2010, *doi: 10.1093/bioinformatics/btq538*.
3. Shlomi et al., *Nat Biotechnol*, 2008, doi: 10.1038/nbt.1487.

BIOMEDICAL TEXT MINING FOR DISEASE GENE DISCOVERY

Sarah ElShal[1, 2,*], *Jesse Davis*[3] & *Yves Moreau*[1, 2].

Depts. of Electrical Engineering (ESAT) STADIUS Center for Dynamical Systems, Signal Processing and Data Analytics[1], *iMinds Future Health*[2] *and Computer Science (DTAI)* [3], *KU Leuven.* [*]*sarah.elshal@esat.kuleuven.be*

How often do you type a question on Google and get back your answer right away? A few decades ago people could not imagine that most information could only be a keystroke or mouse-click away. We are currently facing a phase of digital revolution with lots of opportunities that if handled successfully, can take us to an era we never imagined. In this work we focus on one such opportunity, which is biomedical text mining. We are developing a Google-like tool that given any disease query, the tool returns back an ordered list of the genes most related to that disease. We evaluate our tool based on the disease-gene associations in the OMIM catalog[1]. Our tool has a recall of 73% and 86% in the top 10 and 100 returned genes respectively.

INTRODUCTION

PubMed comprises more than 23 million citations for biomedical literature from MEDLINE, life science journals, and online books[2]. This huge quantity of electronic literature makes it challenging for biologists to search the PubMed corpus for any desired information beyond simple text retrieval. There exist several tools that further identify biomedical concepts given any piece of text such as GoPubMed[3], and others that further apply rule-based strategies to relate biomedical concepts to each other such as BITOLA[4]. In this work we apply text mining techniques to identify and discover links between diseases and genes based on the biomedical literature.

METHODS

For each gene recorded in "Entrez Gene", we record the list of PubMed abstracts linked to it as indicated on GeneRIF[5]. Then for each disease query, we record its list of abstracts as retuned by the online query system of PubMed. To measure the strength of the association between a disease and a given gene we investigate **three approaches**.

The first one is the **explicit approach** where we rely on the clear association between the disease and the gene as reported in the literature. The higher the count of abstracts that mention both of them, the stronger the association is. We use **Fisher's exact test** to evaluate such association. Hence we measure how significant this count of common abstracts is, relative to 1) the raw count of abstracts that mention each of them separately and 2) the total count of abstracts on PubMed.

The second one is the **implicit approach** where we go a further step and try to discover hidden associations between the disease and the given gene. Given the list of abstracts linked to a disease/gene, we generate a term profile that defines this disease/gene. The higher the count of shared terms between the disease and the gene profiles, the stronger the association is. Here, we apply the **cosine similarity measure** to decide how significant this count of shared terms is, relative to the total count of terms that define the disease and the gene separately.

The third one is the **combined approach** where we consider the two association signals coming from the explicit and the implicit approaches. For a given gene, we select the stronger signal to account for its association to the given disease.

RESULTS & DISCUSSION

OMIM provides us with a list of disease-gene annotations [~6000 diseases] based on experimental evidence. We refine this list in 4 consecutive steps:

1- We remove non-confirmed annotations.
2- We discard diseases which have very few (<10) PubMed abstracts linked to them.
3- We normalize gene symbols such that we map all gene synonyms to the official gene symbol.
4- We combine diseases which refer to the same disease concept together.

Hence, we have a refined list of disease-gene annotations [~1200 diseases] based on OMIM. For every disease in this list, we rank the human genome according to how strong every gene is associated to the given disease based on our predefined approaches. Then we count how many OMIM-genes appear in our top 10, 25, 50, and 100 ranked-genes for that disease. We then calculate the average recall over all the disease queries found in our modified OMIM list.

We present the performance of each approach in Table 1. We observe that the explicit approach has an average recall of 68% in the top 10 vs. 78% in the top 100. Similarly, the implicit approach has an average recall of 65% in the top 10 vs. 79% in the top 100. We also observe that selecting the stronger signal via the combined approach achieves the best recall of 73% in the top 10 and 86% in the top 100.

Table 1 The average recall in the top 10, 25, 50, and 100.

	Top 10	Top 25	Top 50	Top 100
Explicit	68%	73%	76%	78%
Implicit	65%	71%	75%	79%
Combined	**73%**	**79%**	**83%**	**86%**

These results show the potential for text mining to discover links between diseases and genes in the biomedical literature. Since a high percentage of the top ranking links is already experimentally-validated, we can highlight the other percentage as potential candidates for further validation.

REFERENCES

1. http://omim.org/
2. http://www.ncbi.nlm.nih.gov/pubmed/
3. http://www.gopubmed.com
4. http://ibmi.mf.uni-lj.si/bitola/
5. http://www.ncbi.nlm.nih.gov/gene/about-generif

BIOINFORMATICS AND SYSTEMS BIOLOGY MASTERS: BRIDGING THE GAP BETWEEN HETEROGENEOUS STUDENT BACKGROUNDS

S. Abeln[1,2,] D. Molenaar [1,3] K.A. Feenstra[1,2,5] H.C.J. Hoefsloot [4] B. Teusink [1,2,3] and J. Heringa[1,2,3,5].*
*VU University Amsterdam: [1] IBIVU Centre for Integrative Bioinformatics [2] Dept. of Computer Sci. [3] Dept. of Molec. Physiology [4] University of Amsterdam, Swammerdam Institute for Life Sciences [5] NBIC Netherlands Bioinformatics Center, Nijmegen. *s.abeln@vu.nl*

Teaching students with very diverse backgrounds can be extremely challenging. We use the Bioinformatics and Systems Biology MSc in Amsterdam as a case study to describe how the knowledge gap between heterogeneous backgrounds can be bridged. A mix in backgrounds can be turned into an advantage by creating a stimulating learning environment. Mixing students from different backgrounds in a group to solve a complex task, creates an opportunity for the students to reflect on their own abilities. We explain how a truly interdisciplinary approach to teaching helps students of all backgrounds to achieve the MSc end terms.

INTRODUCTION

Designing a course or curriculum for students from different backgrounds can be extremely challenging. Typical challenges include creating assignments with enough complexity to reach the right academic level, avoiding the pitfall of teaching a separate course for each background and motivating students to work on gaps in their knowledge. Below we use the MSc Bioinformatics and Systems Biology in Amsterdam as an example to show how such problems may be overcome. The Figure 1 shows the diversity in background of the students entering the programme. The current year (2013/2014) has seen about 40 new students coming from no less than 34 distinct bachelor programmes!

OVERCOMING INITIAL DIFFERENCES

The programme has a very intensive start with two compulsory courses, on bioinformatics and on systems biology. For the first two months, students have 40 hours of total workload, including 24 contact hours per week which are divided up between lectures, assignment classes and conversion classes. This sets the pace for the rest of the programme. It also means that students get to know each other well; this tends to help them through the rest of the curriculum.

The three different conversion classes are provided in the topics of Molecular Biology, Mathematics and Programming , in which we try to give students the means to start working on their deficiency independently. For some students, it is necessary to take further courses to follow their desired profile, typically Molecular Biology or Programming (Java/Python) for the Bioinformatics profile, and Mathematics or Molecular Biology for the Systems Biology profile. It is very important to note that programming and good mathematical skills may, to our observations, indeed be acquired after the BSc level– contrary to popular belief within those disciplines.

CURRICULUM OUTLINE

After these two months of parallel compulsory courses, the students can choose either a Bioinformatics or Systems Biology profile. The profile consists of three courses specially designed for the profile, and three optional courses. Good students with few deficiencies are allowed to follow both profiles, by choosing their optional courses strategically. Other students may need to use these optional courses to become more proficient in programming, mathematics or molecular biology.

After a year of courses, students go on to do independent research projects in their second year, typically one larger (major) project, and one smaller (minor) project. The major project needs to lie within the field of the chosen profile.

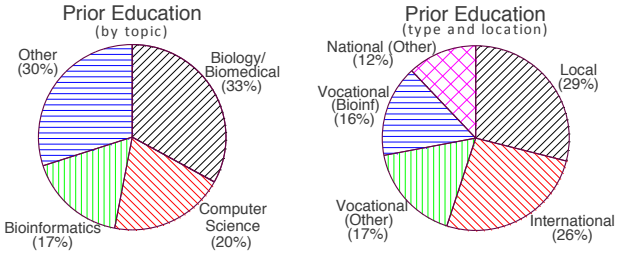


***FIGURE 1**. Heterogeneity of student background in topic (left), type – vocational or scientific – and location (right).*

JOB PROSPECTS

With either strong mathematical or programming skills, and the ability to work with people from a wide variety of backgrounds, our students do not have much trouble finding suitable employment. About one third of students find jobs outside of academia and outside of bioinformatics or systems biology, mostly in data analysis and/or IT.

More than half of the students go on to do a PhD; for students with a vocational BSc, perhaps surprisingly, this fraction also holds. Topics of PhD projects have a very broad range, just like the internship projects, including theoretical studies using mathematical modelling, combinations of experiment with data analysis and the development of new algorithms.

LEARNING POINTS

From a student's perspective, working with peers from different backgrounds is an invaluable experience, and highly regarded in every field of employment. The intensive start setts the pace and helps establishing basic knowledge. Students can be flexible in choosing a curriculum according to their interests. Finally, during all courses and project work, our focus lies on the true interdisciplinarity of bioinformatics and systems biology

JQCML: A JAVA API FOR QUALITY CONTROL FOR MASS SPECTROMETRY EXPERIMENTS

Wout Bittremieux[1,2]*, *Pieter Kelchtermans*[3,4,5], *Dirk Valkenborg*[5,6,7], *Lennart Martens*[3,4] & *Kris Laukens*[1,2].

Dept. of Mathematics and Computer Science, University of Antwerp[1] ; *Biomedical Informatics Research Center Antwerp (biomina), University of Antwerp / Antwerp University Hospital*[2] ; *Dept. of Medical Protein Research, VIB*[3] ; *Dept. of Biochemistry, Faculty of Medicine and Health Sciences, Ghent University*[4] ; *Flemish Institute for Technological Research (VITO)*[5], *I-Biostat, Hasselt University*[6], *CFP-CeProMa, University of Antwerp*[7]. **wout.bittremieux@uantwerpen.be*

QcML is a new approach towards a standardized format for quality control metrics for mass spectrometry experiments. We here present jqcML, an open-source Java API for working with qcML data.

INTRODUCTION

In order to provide a pervasive and standardized means to report quality control information for mass spectrometry experiments, the qcML standard[1] has recently been developed. The qcML standard aims to support an automated quality control pipeline by providing a set of useful metrics that can be calculated on the acquired data. Additionally, it aims to provide a standard format for the exchange of these metrics.

To exchange qcML data an XML-based file format has been developed. This is a universal format that captures metrics and metadata about all kinds of mass spectrometry experiments. As such, the qcML file format can be used as a container to separate quality control information from the actual data analysis.

Here we present jqcML, a fully operative Java API for working with qcML data. Firstly, jqcML provides a complete object model to interpret and manipulate qcML data, while retaining a small memory footprint without sacrificing the overall speed of data access. Furthermore, jqcML is able to interact with both XML-based qcML files and a qcDB relational database. This interaction is abstracted, so the user is able to work with both sources of qcML data in a consistent way.

METHODS

The main approach to exchange qcML data will be through the XML-based qcML files. To handle this approach, jqcML is able to operate on qcML files by reading a full qcML file or only a specific part of a qcML file, and by creating and writing qcML files.

In order to perform input and output on qcML files, the Java Architecture for XML Binding (JAXB) is used. By annotating specific elements of the object model, a mapping between the object model and the XML structure defined by the XML schema is constructed. This allows a translation from qcML files to the object model, and vice-versa. Using JAXB we are able to both read and write qcML files, thus enabling the interpretation of existing qcML files, as well as the creation of new qcML files.

When interpreting data from a qcML file, special care is taken to be able to manipulate files that are arbitrarily large. It is possible to only read a specific part of a file by using an XML indexer component. This prevents having to read the full, possibly (very) large, qcML file into memory, while still being able to retrieve the required content.

Besides the XML-based file format, qcML data can also be stored in a relational database, called qcDB. Equivalent to the XML-based file format, jqcML provides an application layer to be able to read and write qcML data from a qcDB. This interaction is abstracted, enabling the user to interface with an XML-based file or a qcDB in a consistent way. Consequently, jqcML can also be used as a converter between the XML-based file format and a qcDB, and vice-versa.

RESULTS & DISCUSSION

OpenMS[2], an open-source library for LC/MS data management and analyses, provides a (modular) tool to calculate qcML data. By using this pipeline raw files detailing mass spectrometry experiments can easily be processed to output a qcML file.

Using this data it is possible to perform advanced analyses between different runs. By performing a large-scale analysis on specific quality metrics, it is possible to obtain a classification between experiments. Subsequently, based on this classification quality boundaries can be determined. Finally, using these boundaries, thresholds can be defined in order to flag bad experiments.

The qcML standard will be finalized by the end of the year. Using libraries such as jqcML, the capabilities of the qcML standard can easily be harnessed. This will enable the users to provide easier and better quality control for their mass spectrometry experiments, resulting in more high-quality results.

REFERENCES

1. qcml – A XML format for quality related data of mass spectrometry instruments. https://code.google.com/p/qcml/ (2013).
2. Kohlbacher, O. *et al.* TOPP--The OpenMS Proteomics Pipeline. *Bioinformatics* **23,** e191–e197 (2007).

IDENTIFYING INTERESTING FREQUENT PATTERNS IN COMPLEX BIOLOGICAL DATA WITH MIME

Stefan Naulaerts[1,2], *Sandy Moens*[1], *Pieter Meysman*[1,2], *Wim Vanden Berghe*[3], *Bart Goethals*[1] & *Kris Laukens**[1,2].

Department of Mathematics and Computer Science, University of Antwerp[1] ; *Biomedical informatics research center Antwerpen (biomina), University of Antwerp*[2] ; *Department of Biomedical Sciences, University of Antwerp*[3].

**kris.laukens@uantwerpen.be*

With rapid accumulation of complex biological datasets, the need for specialized techniques to rapidly identify patterns remains growing. One powerful technique for the detection of hidden motifs is frequent itemset mining, which has been an integral part of various bioinformatics workflows. However, there are various practical obstacles to the use of frequent itemset mining, such as lack of user-friendliness, pattern redundancy and lack of user feedback in the mining process. As an answer to these shortcomings we present the MIME software framework, together with several real-life biological use cases. Using MIME, we successfully identified sets of interesting patterns that could be validated.

INTRODUCTION

More and larger datasets are continuously being released due to the omnipresence of high-throughput methods. Although this is a powerful driver for systems biology research, the data scale makes discovery and analysis of patterns in the data challenging. Since their first introduction, frequent item set mining has proven especially useful in capturing characteristics of a datasets and the algorithms are growing towards succinct summarization of datasets into the relevant parts.

Although promising for bioinformatics research, the information about this set of techniques is very scattered. More often than not, the implementations are only available as command line tools that are not user-friendly and need to be compiled from source. Results are also presented as a static list of candidate patterns, with no room for the user to interact and refine the output. Furthermore, what constitutes an *interesting* pattern is hard to define objectively. Current criteria to define *interestingness*, are based on computational aspects that completely ignore the underlying biological processes and thus may not be ideal.

In response to these shortcomings, MIME [1] was developed and applied to several biological use cases, featuring very diverse data types. MIME or "Making Interactive Mining Easy" is a software framework that incorporates various frequent itemset mining and post-processing methods that have been thoroughly described and validated in literature. However, the strength of these tools increases dramatically when they are combined into an iterative mining workflow. MIME provides a graphical interface through which the user can rapidly expand or filter his results and as such, interact with his data to find core aspects relevant for his research in a transparent manner.

METHODS

Using MIME, we investigated several use cases for frequent itemset mining in biological datasets. To this end, we used popular mining algorithms that are included in the software, including variants of Apriori [2].

We investigated the co-occurrence of protein domains within and between proteins using InterPro [3] and the IntAct [4] protein-protein interaction network. We also conducted a functional analysis using the Gene Ontology [5] and searched for motifs in gene expression profiles obtained from Colombos [6].

RESULTS & DISCUSSION

For each of the datasets we used in our case study, MIME and its mining algorithms were able to extract biologically relevant patterns, which we could validate in various ways. We were able to sufficiently filter the initially large amount of itemsets and generate appropriate association rules that can be used for classification.

Overall, MIME should provide an easy-to-use environment for life scientists, to allow them deeper exploration of complex datasets. The software is unique in the sense that it allows rapid iterative mining and directly incorporates the user feedback in the mining process.

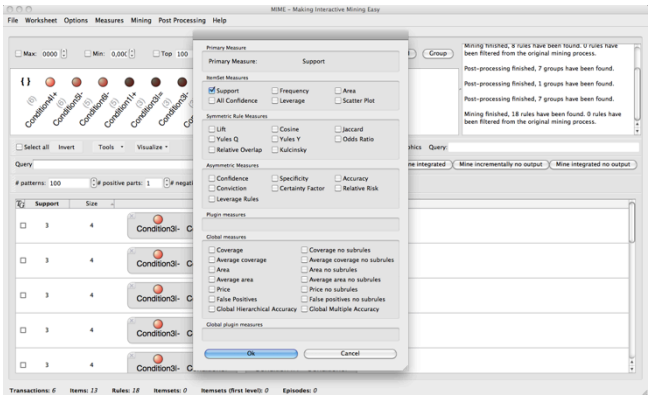


***FIGURE 1**. Overview of some of the "classic" implemented quality measures in MIME. The background shows elements of a toy dataset (left upper pane) and brief summary of the mining output (right upper). The bottom panel shows the actual itemsets with the active measures.*

REFERENCES

1. Goethals B et al. ECML PKDD'11. **3**:634–637 (2011).
2. Agrawal R et al. ACM SIGMOD. **22**:207–216 (1993).
3. Hunter S et al. Nucl Acids Res. **40**:D306–D312 (2012).
4. Kerrien S et al. Nucl Acids Res. **40**: D841-D846 (2012).
5. Ashburner M et al. Nature Genet. **25**: 25-29 (2000).
6. Engelen K et al. PLoS One. **6**: e20938 (2011).

FACILITATING COMPUTATIONAL BIOLOGY AND BIOINFORMATICS ON HPC SYSTEMS USING EASYBUILD

Fotis Georgatos[1,*], Nils Christian[1], Kenneth Hoste[2], Cedric Laczny[1], Thekla Loizou[3], Wiktor Jurkowski[1], Andreas Panteli[3], Reinhard Schneider[1], George Tsouloupas[3], Jens Timmerman[2], Vasilis J. Promponas[4]

*Luxembourg Centre of Systems Biomedicine/University of Luxembourg[1] , Ghent University[2], The Cyprus Institute[3] , Bioinformatics Research Laboratory/University of Cyprus[4] , *Fotis.Georgatos@uni.lu*

Environment modules constitute a best practice approach across HPC sites to manage the software complexity of multi-user environments. Computational Biology and Bioinformatics is a domain characterized by heterogeneous software resources, constant growth and need for adaptation to emerging technologies, and a diverse spectrum of end-users. We demonstrate an approach for fully automated software installation eliminating common bottlenecks, allowing sharing of software recipes; this is achieved in a scriptable manner using an open software framework, named EasyBuild.

INTRODUCTION

Modern bioinformatics research relies upon a large array of heterogeneous software tools developed and maintained by diverse teams working under different environments and objectives. High Performance Computing (HPC) centers are exploited by growing communities of bioinformaticians and computational biologists, therefore the complexity presented when delivering the related software stacks is growing. To overcome this bottleneck, we rely on the recently established EasyBuild [1, 2] platform, to develop recipes for packages, offering ready-made solutions in an open source repository for immediate access by domain experts. Even more so, modern systems –even of small size– present similar parallelization capabilities, such as GPUs and multi-core CPUs, which certainly pose similar challenges. The proposed technique has already been practiced across several HPC sites. The open source repository is available online [3]. We anticipate that this paradigm shift in HPC software reusability will enhance the community's productivity.

METHODS

Environment modules [4] are a de facto standard on HPC systems, because it allows dealing with the diversity of software builds and versions in an elegant and well-defined manner. Also, modules are usable for dependency resolution, exposing the software stack in a transparent manner.

EasyBuild offers a collection of desirable features, making it a generic platform suitable for reproducibility of computing environments for various scientific domains. Specifically, it permits independently installing multiple versions of software packages, taking care of all dependencies or multiple compilers, libraries and tools, also deployed using EasyBuild. Importantly, all these are achieved through simple Python-based configuration files:

- *easyconfigs* (.eb files) provide a concise description of the basic parameters for builds/installations which follow a standard, eg. configure/make/make install
- *easyblocks* (.py files) are Python modules that are required whenever a custom build/install procedure needs to be followed.

Also, with the proposed method, a record of the installation logs is kept for future reference and debugging purposes, and it is also possible to keep track of installation configuration in a version control system (e.g. git).

A key benefit of this approach, in addition to reusability and ease of use, is the reproducibility and deterministic outcomes of the build process. The precise build process for a software package compiled through EasyBuild on a specific system can be expected to be reproducible on another system, which is extremely useful for scientific workflows.

RESULTS & DISCUSSION

The proposed scheme has at this point found its way into the production environment of HPC sites across a handful of countries (the authors' institutions plus Gregor Mendel Institute, Austria). In Bioinformatics and Computational Biology, we already have 60 packages in reproducible build configurations. Among them the following ones stand out:

- Generic libraries /packages (BioPerl, BioPython)
- Sequence Analysis tools (BLAST/NCBI_Toolkit, EMBOSS, FASTA, HMMER, mpiBLAST)
- Phylogenetic Analysis tools (MrBayes, RAxML)
- Next Generation Sequencing tools (ABySS, BWA, Bowtie, TopHat, Velvet)
- Other popular tools (e.g. Rosetta, GROMACS)

More related packages are available, such as Bioconductor (under the R framework), MATLAB Bioinformatics Toolbox (under the respective Matlab module).

Based on our experience, delivering a structured software environment on scientific computing platforms has now become a tractable task. However, this is achievable provided that the right mix of tools is employed. EasyBuild is a tool facilitating the collaboration between HPC sites, supported by an open and enthusiastic community. Contributions of any form (feedback, code, new ideas) are much appreciated.

REFERENCES

1. Hoste K. et al. *EasyBuild: Building Software With Ease*; November 2012 @ PyHPC-2012 workshop at Supercomputer 2012 conference.
2. http://hpcugent.github.io/easybuild/
3. https://github.com/hpcugent/easybuild/wiki/List-of-supported-software-packages, accessed 20/10/2013.
4. Furlani JL. In Proceedings of 5th Large Installation Systems Administration Conference (LISA V), pp. 141-152, San Diego, CA, September 30 - October 3, 1991.

NGS LOGISTICS: DATA INFRASTRUCTURE FOR EFFICIENT ANALYSIS OF NGS SEQUENCE VARIANTS

Amin Ardeshirdavani[1,*], *Erika Souche*[2], *Luc Dehasbe*[3], *Jeroen Van Houdt*[3], *Joris Vermeesch*[3], *Yves Moreau*[1].

KU Leuven ESAT-STADIUS Center for Dynamical Systems, Signal Processing and Data Analytic – iMinds Future Health Department [1], KU Leuven Laboratory for Molecular Diagnosis [2], KU Leuven Department of Human Genetics (Genomics Core) [3]. [*]amin.ardeshirdavani@esat.kuleuven.be

Next-Generation Sequencing (NGS) is quickly becoming a key tool in research and diagnostics of human Mendelian, oligogenic, and complex disorders [1]. Currently, there are about 2,560 sequencers located in 920 centers all around the world [2]. The average turnaround time for exome sequencing is four weeks, and eight weeks for complete genome sequencing. Because the price and turnaround time for sequencing has dramatically decreased over the past decade, large amounts of human sequencing data are now available. It is estimated that between 50,000 and 100,000 individuals have been sequenced worldwide by now. This raises major challenges in terms of data storage, management, exchange, and for federated analysis. This also raises substantial ethical and privacy issues.

INTRODUCTION

Estimates for data storage reveal that at least 100 GB are necessary to store all files (fastq, bam, vcf, etc.) related to the whole genome sequence of one individual [3]. To process the large-scale data generated by human genome sequencing, well-designed infrastructures supported by powerful computational resources and large storage facilities are essential. When performing case studies or association studies, lack of reference/control cases or limitation in the number of patient cases is a major bottleneck. To resolve this, we need to expand the study by finding additional related cases. This will be only possible with effective data exchange or collaborative querying systems. Furthermore, downloading data from public data repositories is time consuming and not cost effective. Besides, privacy and security issues restrict the effectiveness of the collaboration between different research institutes.

METHODS

We have developed an online tool (NGS-Logistics) that fulfills all requirements of a successful application that can process data inclusively and comprehensively from multiple sources while guaranteeing privacy and security. NGS-Logistics is a web-based application providing a data structure to analyze NGS data in a distributed way. A key feature is that queries are executed across multiple centers without moving primary data around. In the first step we are focusing on the detection of variation over different kind of phenotypes.

Data is available at several different sequencing centers. For different research groups, their Principal Investigators (PI) are added to the system. Datasets of sequencing samples with full project description and sample information are defined and associated to PIs. Users can request access to the different datasets. Well-defined Access Control List (ACL) mechanisms provide access of users to the right datasets and operations. This guarantees appropriate access control and privacy of the data. Each operation consists of queries and tasks, which we have designed to analyze data. Queries submitted by users are picked up by a job controller, and sent to all centers. Depending on the type of query, desired results are generated and are returned to the main system.

Results are divided into two parts, the first part is related to the samples to which the user has authorized access, and for which the users can see all details. The second part contains results for the whole population, for which the user has only access to some aggregate statistics without details. An example of such a query would be to review the mutations present at a single genomic position in each individual patient from a set of patients to which the user has authorized access (1st part) and to contrast these results with background frequency of mutation in the reference populations (2nd part).

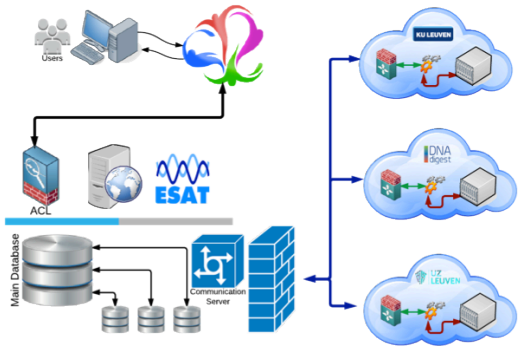


Figure 3: NGS-Logistics workflow

RESULTS & DISCUSSION

The pilot version of NGS-Logistics has been installed and is currently being beta-tested by users at the Center for Human Genetics of the University of Leuven. Currently we have two installations of the system, the first one at the Leuven University Hospitals and the second one at the Flemish Supercomputing Center (VSC). The development of NGS-Logistics has significantly reduced the effort and time needed to evaluate the significance of mutations from full genome sequencing and exome sequencing, in a safe and confidential environment. This platform provides more opportunities for operators who are interested in expanding their queries and further analysis.

REFERENCES

1. Voelkerding KV et al. *Clin Chem* **55**, 641-658 (2009).
2. Next Generation Genomics: World Map of High-throughput Sequencers [http://omicsmaps.com/]
3. Kahn SD. *Science* **331**, 728-729 (2011).

DBXP: INVESTIGATING THE FUTURE OF INTEGRATIVE BIOINFORMATICS RESEARCH INFRASTRUCTURES IN EUROPE

Lars Eijssen[1,], Jildau Bouwman[2], Anwesha Dutta[1], Nuno Nunes[1], Marijana Radonjic[2], Thomas Kelder[2], Varshna Goelela[3], Stan Gaj[4], Maarten Coonen[4], Michiel Adriaens[5], Wibowo Arindrarto[6], Philip de Groot[7], Magali Jaillard[8], Ben van Ommen[2], Chris Evelo[1].*

Dept. of Bioinformatics-BiGCaT, Maastricht University[1] ; Microbiology and Systems Biology, TNO, Zeist[2] ; Galapagos BV, Leiden[3] ; Dept. of Toxicogenomics, Maastricht University[4] ; Dept. of Experimental Cardiology, Heart Failure Research Center, AMC, University of Amsterdam[5] ; Leiden University Medical Center, Leiden[6] ; Nutrition, Metabolism and Genomics Group, Wageningen UR[7] ; BioMérieux, Division of Advanced Technologies, Lyon, France[8].

[*]l.eijssen@maastrichtuniversity.nl

Sharing and integrating multiple (un)published data sets are pivotal to drive progression of research. By integrating existing technological research infrastructures (RIs), each generating different types of data, study outcome can be optimised. Besides exchange, data sharing requires standardisation and data normalisation. Ideally, a study data infrastructure would allow analysis and querying of results of different studies in combination and at several levels of both study description and study outcome.

DBXP

One of the main developments to integrative data processing, storage, and querying is the Phenotype Data-sharing Infrastructure (dbXP, www.dbxp.org), which incorporates existing research infrastructures (RIs). dbXP allows description of studies, samples and assays using a template-based interface with standardised ontologies. It respects data ownership, but promotes data sharing. As shown in Figure 1, the core study-capture module points to raw and processed data and connects to data processing pipelines for multiple technologies[1]. dbXP currently consists of a simple assay module and complex data processing modules, e.g. for metabolomics and metagenomics. Several instances of dbXP are already operational and new ones can be initiated. For testing purposes a special instance of dbXP has been made available: visit test.dbnp.org to get an account. We warmly welcome you to try!

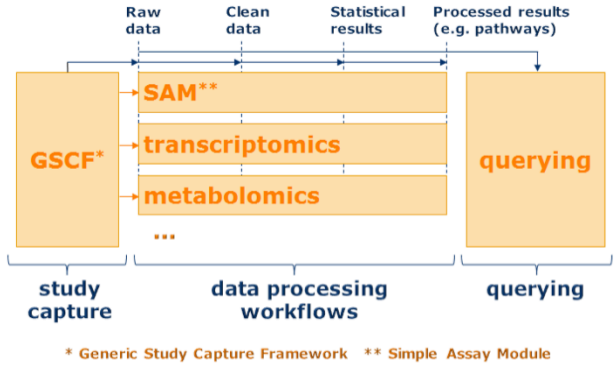


***FIGURE 1**. Schematic representation of the setup of dbXP*

EURODISH

EuroDISH (www.eurodish.eu) is a 3 year FP7 programme with 15 partners, including the worldwide Nutrigenomics Organisation NuGO, focused on identifying existing food and health RIs, their integration, and the need of new RIs that are relevant for innovations in mechanistic research and public health nutrition strategies across Europe. EuroDISH uses the nutritional instance of the dbXP infrastructure, dbNP, for its illustrative case study focused on the metabolic syndrome. Since comparable technologies are often applied in different scientific fields, most of the observations made by EuroDISH will also extrapolate to other fields. Experiences during the EuroDISH survey and case study will be used to advice European research initiatives and to further improve dbXP.

ARRAYANALYSIS

The dbXP infrastructure is extended over time with new data processing workflows, including those available at ArrayAnalysis.org for the handling of omics data types, including quality control, pre-processing, statistics, and pathway analysis[2,3].

The modules can already be used directly at the ArrayAnalysis.org web portal. Automated quality control and normalisation can be performed on data of several types of omics platforms including Affymetrix[2], Illumina (paper in preparation), Agilent and spotted arrays (module and paper in preparation). After normalisation, automated calls can be made to a statistics module, implementing *limma* regression analysis[4]. The outcome from this module, can be forwarded to a pathway analysis module automatically calling PathVisio (www.pathvisio.org)[5]. These calls are made through PathVisioRPC (paper in preparation). For developers, there is a developers site that also has additional beta modules, among which a module for the analysis of Nimblegen methylation experiments. Access will be granted on request.

Also, PathVisio functionality is being extended with other types of analyses, allowing for automated calls of those as well. Pathway, gene, and disease ontology enrichment analyses and storage of the results will allow comparison of studies based on comparable biological outcomes which will be key to the querying mechanisms for dbXP[6].

REFERENCES

1. van Ommen *et al. Genes Nutr* **5**, 189-203 (2010).
2. Eijssen *et al. Nucleic Acids Res* **41**, W71-6 (2013).
3. Norheim *et al. Nutrients* **4**, 1898-944 (2013).
4. Smyth *et al. Stat Appl Genet Mol Biol* **3**, Article 3 (2004)
5. van Iersel *et al. BMC Bioinformatics* **9**, 399 (2008).
6. Evelo *et al. Genes Nutr* **6**, 81-7 (2011).

UPCOMING COURSES *(Maastricht University)*

Microarray Analysis using R and Bioconductor (*January 28-30)*

Molecular Epidemiology of Chronic Diseases (*June 16-20)*

DATA INTEGRATION & STEWARDSHIP CENTRE: TACKLING THE BIG DATA CHALLENGE IN LIFE SCIENCE RESEARCH

Jan-Willem Boiten[1], Jildau Bouwman[2], Bas van Breukelen[3,4], Lars Eijssen[5], Chris Evelo[5], Richard Finkers[6,7], Femke Francissen[8], Celia van Gelder[8,9], Martien Groenen[6], Rob Hooft[8], Ruben Kok[8,], Barend Mons[8,10], Irene Nooren[11], Marco Roos[10], Gabino Sanchez Perez[6,7], René van Schaik[12], Morris Swertz[13].*

*CTMM-TraIT, Amsterdam[1] ; TNO Quality of Life, Zeist[2] ; Utrecht University[3] ; Netherlands Proteomics Centre, Utrecht[4] ; BiGCaT, Maastricht University[5] ; WUR, Wageningen[6] ; Plant Research International, Wageningen[7] ; Netherlands Bioinformatics Centre, Nijmegen[8] ; Radboudumc, Nijmegen[9] ; LUMC, Leiden[10] ; SURFsara, Amsterdam[11] ; eScience Center, Amsterdam[12] ; UMCG, Groningen[13] ; **all authors are active in DTL**. [*]ruben.kok@dtls.nl*

The Dutch Techcentre for Life Sciences (DTL) is collaborative platform of life science and technology research groups in the Dutch clinical & health, nutrition, crop and livestock breeding and industrial microbiology sectors (www.dtls.nl). DTL is established as a public-private partnership of universities, university medical centres, research institutes, life science industry and technology companies (Figure 1). DTL creates an environment that enables Next Generation Life Science research. Collectively, the DTL parties address the major 'big science & big data' challenges in biology R&D that no single organisation can address alone in two interlinked programmes: (i) Data Integration & Stewardship Centre (DISC), and (ii) Experimental Life Science Technologies (LSTECH). This contribution describes the structure of DTL-DISC.

DTL-DISC

DTL-associated scientists and engineers are responsible for data integration & stewardship in various life sciences big data initiatives in clinical and biomedical research in the Netherlands and internationally (e.g. ELIXIR). DTL-DISC is the joint platform that brings together state-of art expertise and resources for crucial steps in the data cycle (http://www.dtls.nl/dtl/programmes/disc.html and Figure 2).

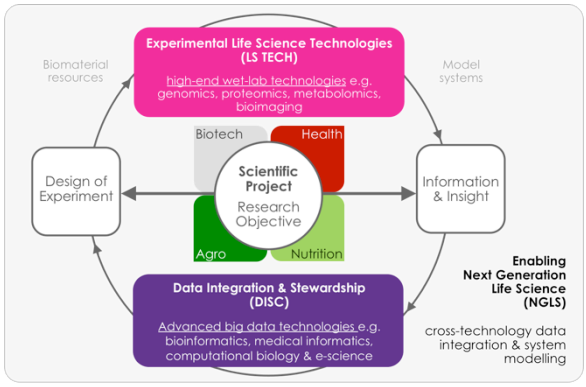
DISC bundles expertise, support and training, as well as tangible tools and infrastructures and forms a national infrastructure that offers access to:

I **Bioinformatics and medical informatics expertise** – professional expertise in design and performance of crucial steps of the data cycle to enhance data integration & stewardship in life science research projects.

II **e-Infrastructure** – High-performance compute, storage & connectivity, incl. cloud services (SURFnet & SURFsara).

III **Tools & Databases** – Engineering of tailor made tools, as well as standardised software platforms & databases, including standardised solutions for interoperability of data sets.

IV **Helpdesk, Training & Education** – easy access to the capabilities of the DTL partnership, and tangible courses for in-depth training and education



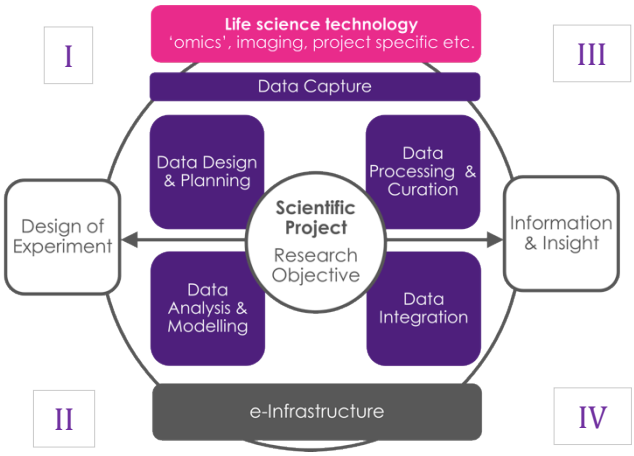


***FIGURE 2**. The data cycle that is reflected in the setup of the DTL-DISC services and infrastructures (I – IV).*

TECHNOLOGY HOTELS

DTL services are organised in the form of technology hotels, including a wide coverage of bioinformatics and systems biology support. More information on all registered technology hotels and how to access them, is available at http://www.dtls.nl/dtl/technology-partners/hotels.html.

'PROJECT AND AREA LIAISONS' (PALS)

DISC will implement mechanisms to get young researchers involved in good data stewardship, following previously successful examples, such as in a large EU project on Systems biology of Microorganisms (SysMO). The idea is that young researchers are rewarded for introducing new ways of working by being provided with extra support for their work and direct influence on the development of these new working methods.

◄ ***FIGURE 1**. Organisation and embedding of the Dutch Techcentre for Life Sciences (DTL)*

BIOINFORMATICS @ DSM BIOTECHNOLOGY CENTER

Léonie Boender-van-Dijk[1,*]
[1]*DSM Biotechnology Center, PO box 1, 2600 MA Delft, The Netherlands.* [*]*leonie.boender-van-dijk@dsm.com*

DSM – Bright Science. Brighter Living.™ is a global science-based company with application areas including food and dietary supplements, personal care, feed, pharmaceuticals, medical devices, automotive, paints, electrical and electronics, life protection, alternative energy and bio-based materials. The DSM Biotechnology Center enables the rapid development of strains for innovative products through classical and rational strain engineering. In this, "Bioinformatics & Modeling" is one of the key enabling competences.

INTRODUCTION

Bioinformatics at DSM encompasses several expertises, including DNA design, Protein design, Pathway and Strain design, Bioinformatics Dbs and algorithms, (Re)sequencing and omics data, and Biostatistics (Figure 1). Three examples concerning these subjects will be addressed.

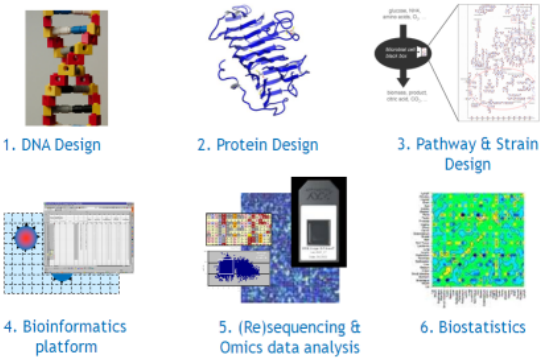


***FIGURE 1**. Bioinformatics @ DSM Biotechnology Center*

DNA DESIGN

Together with Biomax AG, Germany we developed a web interface for DNA construct design[1]. Using a standardized parts repository, novel constructs can be designed and one-click cloning workflows ensure no errors are made during the wet-lab construction. Furthermore, an extensive knowledge management system allows tracking of the experimental results.

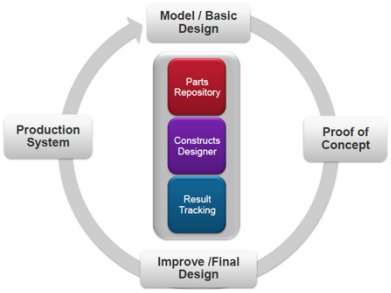


***FIGURE 2**. DNA design*

(RE)SEQUENCING

DSM is a leader in the development of technology for bio-ethanol production from cellulosic feedstocks. The yeast *Saccharomyces cerevisiae* is the organism of choice in the ethanol industry. However, *S. cerevisiae* cannot convert the pentose sugars by nature. These pentose sugars make up a large proportion of the cellulosic hydrolysates. Strains able to ferment both hexose and pentose sugars were acquired through rational strain engineering and classical engineering. Throughout the strain development, strains were sequenced, resulting in a sequenced lineage of fourteen strains. This allowed us to distinguish sequencing errors from important genetic changes and to follow strain development in depth.

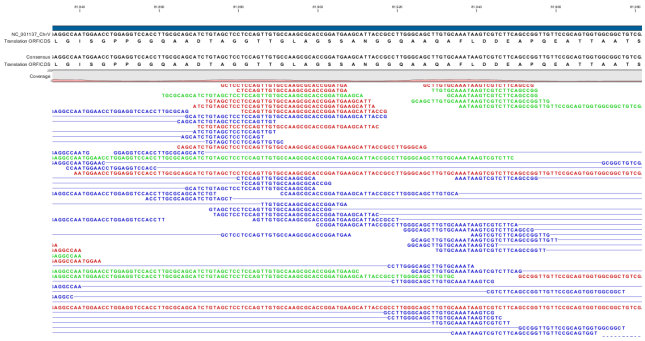


***FIGURE 3**. Alignment of (re)sequencing data*

PATHWAY DESIGN

An example of pathway design is the production of succinic acid from renewable sugars by *Saccharomyces cerevisiae*. Succinic acid can be applied in a wide range of products from food flavor to renewable thermoplastics. A metabolic engineering strategy was employed introducing reductive TCA cycle, glyoxylate shunt and transporter. After proof of principle transcriptomics was applied to generate new leads, which resulted in new improved strains.

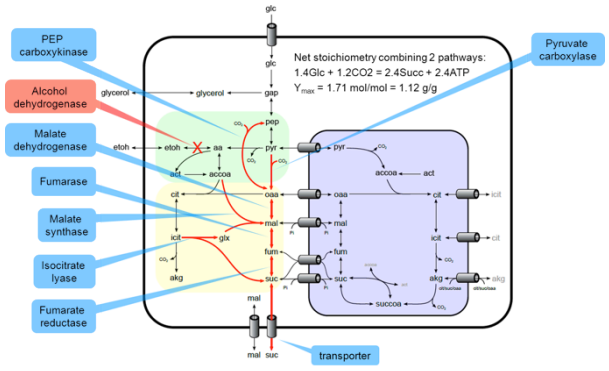


***FIGURE 4**. Pathway design*

REFERENCES

1. Biomax AG: http://www.biomax.com/company/newsletter/archive/Newsletter_Fall_2013.html2.

PACBIO – SINGLE MOLECULE SEQUENCING TO IMPROVE THE NORWAY SPRUCE GENOME ANNOTATION

Yao-Cheng Lin[1], Nicolas Delhomme[2], Stefan Jansson[2], Yves Van de Peer[1] and Nathaniel Street[2]*
*Department of Plant Systems Biology (VIB) and Department of Plant Biotechnology and Bioinformatics (Ghent University)[1]; Umeå Plant Science Centre, Dept. of Plant Physiology, Umeå University[2] *yalin@psb.vib-ugent.be*

We introduce a novel error correction method for Pacific Biosciences cDNA high error rate data. Single molecule transcriptome sequencing provides unambiguous resolution to support the genome assembly and gene structures. The full gene structure can be identified from one continues sequence without assembling short reads.

INTRODUCTION

The recent advances of massive parallel sequencing (RNA-seq) provide great opportunities for gene discovery with unprecedented breadth and depth. The first Norway spruce (*Picea abies*) genome annotation is well supported by RNA-seq data from a broad biological sample collection[1]. However, correct splice alignment of short reads in a complex plant genome such as spruce remains a challenging task due to the presence of huge amounts of repetitive sequences, large gene family sizes and the presence of 'gene-like' fragments. On the other hand, single molecule sequencing instruments can generate multi kilobase sequences with the potential to improve gene discovery without transcriptome assembly. The main drawback of single molecule sequencing however is the high error rate, which hinders current applications in genome sequencing projects. Here, we demonstrate a proof-of-concept long read error correction workflow to improve the read quality in a timely fashion and the application in genome annotation.

METHODS

We generated PacBio cDNA from 14 single molecule real-time (SMRT) sequencing with 1kb and 3kb insert size libraries. High coverage Illumina RNAseq from 22 Norway spruce samples including needles, stems and cones collected at different developmental stages and different time points were obtained from previous study[1]. The in-house MegaBLASTN pipeline and LSC[2] (long read error correction tool) were used to correct the PacBio sequencing error based on the Illumina data (Figure 1).

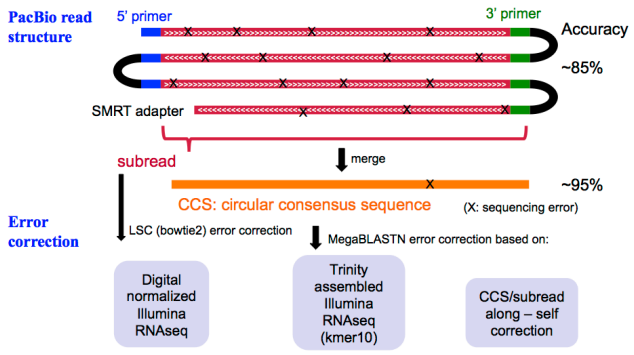


FIGURE 1. *PacBio read structure and data set used in error correction*

RESULTS & DISCUSSION

Long reads allow the well developed long read splice aligner such as GMAP to precisely determine splice site junctions. Error corrected reads improve the mapping rate to the genome sequence. The in-house MegaBLASTN strategy outperforms the existing tool (Figure2).

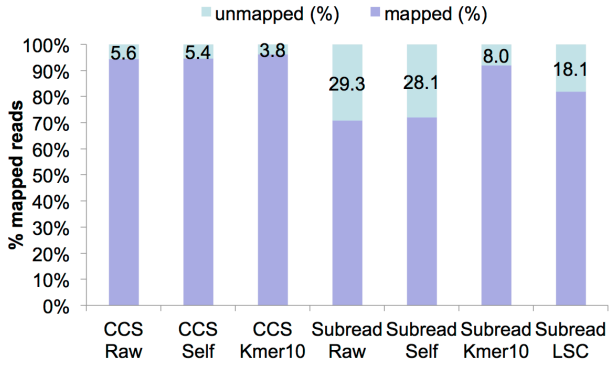


Figure 2*. Performance comparison of error correction strategies. (Self: PacBio reads were corrected by MegaBLASTN based on the same input PacBio reads. Kmer10: PacBio reads were corrected by MegaBLASTN based on the Illumina RNAseq assembly (Kmer10). LSC: PacBio reads were corrected by the LSC pipeline.)*

More than 40% of the high confidence genes were supported by transcriptome evidence. The relatively lower number of supported genes in the PacBio data is due to the low complexity of the input material. That is, current cDNA normalization step is not sufficient to reduce the highly expressed genes (Table 1).

	PacBio	Kmer10	454 assembly
Genes (% of total genes)	11,757 (41.5%)	16,489 (58.1%)	12,712 (44.8%)
Exons (% of total exons)	14,403 (16.8%)	22,873 (26.7%)	20,561 (24.0%)

TABLE 1*: Summary of supported gene models based on different transcriptome evidence.*

In this study, long reads provide direct evidence to support the RNA-scaffolded genome sequence in the first release of the assembly. 4,195 of the RNA scaffolded gaps were supported by PacBio reads and >3,300 group of scaffolds can be further extended/scaffolded by PacBio reads. Furthermore, long introns can be easily verified by long reads without laborious PCR confirmation.

REFERENCES

1. Nystedt B *et al. Nature* ***497***, 579-584 (2013).
2. Au KF *et al. PLoS ONE* **7**(10): e46679 (2012).

A RANDOM FORESTS BASED BREAST CANCER DIAGNOSIS TOOL USING CIRCULATING MIRNA EXPRESSION

Stephane Wenric[1,2*], *Pierre Freres*[2], *Claire Josse*[1,2], *Vincent Bours*[1], *Guy Jerusalem*[2].
University of Liege, GIGA-Research, Human Genetics Unit[1]*; University of Liege Hospital (ULg CHU), Medical Oncology Laboratory*[2]. *s.wenric@ulg.ac.be*

Breast cancer is the leading cause of death by cancer among women and there is a need to improve diagnosis methods. MicroRNAs (miRNAs) are noncondig RNAs that regulate gene expression and many have been implicated in breast cancer. Here, we show that an accurate diagnostic tool for breast cancer can be built based on the expression levels of 8 circulating miRNAs (out of 188 probed miRNAs) and the use of the Random forests classification algorithm.

INTRODUCTION

An efficient classification model has been developed using circulating miRNAs expression levels as features in an ensemble tree-based supervised learning algorithm (Random forests[1]). This model has been validated on an independent cohort.

METHODS

The expression levels of 188 circulating miRNAs was determined for 101 patients with breast cancer and 20 controls.

The individuals were randomly separated into 2 independent cohorts with the same patients/controls ratio:

- profiling cohort, *n = 85*
- validation cohort, *n = 36*

A Random forests model using all 188 miRNAs has been built on the profiling cohort, yielding two variable importance metrics (*mean decrease in accuracy* and *mean decrease in Gini*[2])

Based on these metrics, the miRNAs were ranked, and a selection of 15 miRNAs (which were all ranked among the 20 first miRNAs of both rankings) has been performed.

From these 15 miRNAs, 32767 combinations of 2 to 15 miRNAs have been generated.

A Random forests model was then built for each of these combinations, and its classification performance was assessed by carrying ten-fold cross-validation and comparing the resulting AUC.

Finally, the model built with the profiling cohort and the best performing combination of miRNAs has been validated by predicting the classes of each individual of the validation cohort.

RESULTS & DISCUSSION

The best performing combination of miRNAS among the 32767 combinations was composed of 8 miRNAs and yielded an AUC of 0.9625 when using ten-fold cross-validation on the profiling cohort.

The model built with the profiling cohort and said combination of miRNAs has been validated by predicting classes for the independent validation cohort, and gave an **AUC of 0.9521875.**

To our knowledge, this is the first time the Random forests method is used to perform classification using circulating miRNAs expression levels as features.

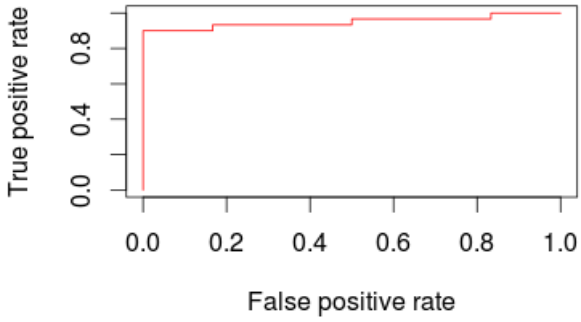


***FIGURE 1**. ROC curve obtained through validation of the final model (built on the profiling cohort, with 8 miRNAs) on the independent validation cohort. AUC = 0.952*

REFERENCES

1. Breiman L. "Random forests." *Machine learning* **45**, 5-32 (2001).

2. Liaw A & Wiener M. Classification and Regression by randomForest. *R News* **2**, 18-22 (2002).

REGRESSION WITH ENRICHED RANDOM FOREST

Martin Otava[1,*], *Ziv Shkedy*[1], *Dhammika Amaratunga*[2], *Javier Cabrera*[3] & *Yung-Seop Lee*[4].
Interuniversity Institute for Biostatistics and Statistical Bioinformatics, CenStat, Universiteit Hasselt[1]; *Department of Nonclinical Biostatistics, Janssen Research & Development, LLC, USA*[2]; *Department of Statistics, Rutgers University, USA*[3]; *Department of Statistics, Dongguk University, South Korea*[4]. *martin.otava@uhasselt.be*

Enriched random forest (ERF) is the modification of the random forest algorithm designed to overcome problems arising in case of extensive number of predictors completely unrelated with response. Weights added into sampling process increase frequency of picking the informative predictors and so prevent underestimation of their relationship with response. Classification ERF weights can be simply computed using two (or more) groups comparison methods, but selection of proper weights in case of both continuous response and continuous predictors does not have any clear solution. We explored several possibilities to choose weights for ERF. All the methods are available in R package ERF.

INTRODUCTION

Random forest[1] is well established method and proved itself to be a useful tool that produces reliable results under various scenarios. One of its key features is the random sampling of predictors to be evaluated within each particular node. However, when most of the predictors are non-informative, we are sampling mostly irrelevant predictors and build trees without any gained value. Enriched random forest (ERF) addresses this issue by adding the weights into sampling process. Its core feature is the way how the weights are defined. The classification case was successfully addressed before[2] and in this contribution we will focus on the regression case.

METHODS

Backbone of ERF is a classical random forest. We grow multiple trees and aggregate their results at the end. For each tree, samples are divided into "in-bag" and "out-of-bag" sets. While former is used as a training sample for tree growing, latter serves as evaluation set and is used for prediction at final step. When a tree is constructed, only a subset of predictors is considered as candidates for split variable in each node. In classical random forest, this sampling is completely random. We replace it with the weighted sampling, when weights are computed for each tree separately, since only in-bag samples are taken into account. There are multiple options how to define weights, depending on the type of relationship we would like to follow. We consider weights based on:

- Pearson correlation p-values,
- Hoeffding's D p-values,
- absolute value of Pearson correlation,
- R^2 of simple tree.

Third and fourth methods produce weights in a straightforward manner, they are proportional to obtained quantities. First and second methods are based on p-values, but before converting them into weights, we transform them into q-values[3] to suppress overfitting. Advantage of q-values is that they would give same weights to all genes when there is no relationship. In the same setting, P-values follow uniform distribution, hence, some of them will be small by chance.

Each tree produces prediction for all its out-of-bag samples. Final prediction is an average of all these predictions. The selection of interesting predictors is based on number of trees where the predictor was selected as the splitting variable.

RESULTS & DISCUSSION

We demonstrated on simulated data sets (Figure 1) and several microarray case studies with large number of genes that the enrichment improves quality of both prediction and feature selection. Further investigation of properties of ERF is subject of follow up research.

Alternative approach is filtering of genes before running random forest itself. Disadvantage of such a method is dichotomy of decision while in ERF we keep all predictors, only varying their prior importance.

ERF is easy to implement and efficient solution in case of large amount of predictors and only small fraction of truly interesting features among them. Additional weighting methods can be considered, depending on a relationship of interest between the response and predictors .

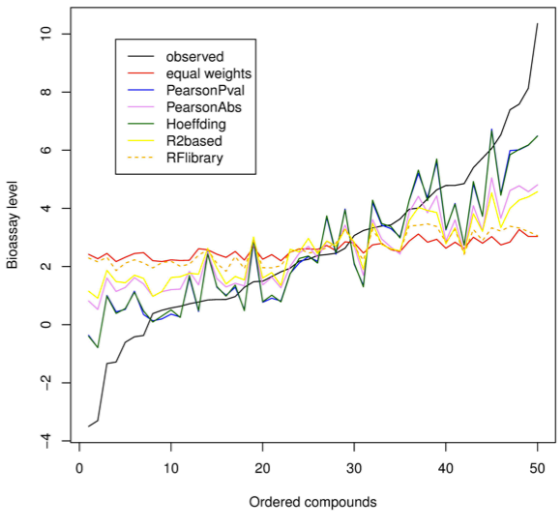


***FIGURE 1**. Example. Improvement in prediction while using different weights when compared to standard random forest algorithm.*

REFERENCES

1. Breiman L. *Mach Learn* **45**, 5-32 (2001).
2. Amaratunga D, Cabrera J & Lee YS. *Bioinformatics* **24**, 2010-2014 (2008).
3. Storey JD & Tibshirani R. *Proc Natl Acad Sci* **100**, 9440–9445 (2007).

TRANSPOSABLE ELEMENT ANNOTATION USING RELATIONAL RANDOM FORESTS

Eduardo P Costa[1], Leander Schietgat[1,], Ricardo Cerri[2], Celine Vens[1,3,4], Carlos N Fischer[5], Claudia M A Carareto[6], Jan Ramon[1] & Hendrik Blockeel[1,7].*

Department of Computer Science, KU Leuven[1]; Instituto de Ciências Matemáticas e de Computação, Universidade de São Paulo, Brazil[2]; VIB Inflammation Research Center, Ghent[3]; Department of Respiratory Medicine, Ghent University[4]; Department of Statistics, Applied Mathematics, and Computer Science[5], and Department of Biology[6], UNESP São Paulo State University, Brazil; Leiden Institute of Advanced Computer Science[7]. [*]Leander.Schietgat@cs.kuleuven.be

Transposable elements (TEs) are DNA sequences that can change their location within the genome. They contribute to genetic diversity within and across species and their transposing mechanisms may also affect the functionality of genes. Accurate annotation of TEs is an important step towards understanding their effects on genes and their role in genome evolution. We present a framework for annotating TEs which is based on relational random forests. It allows to naturally represent the structured data and biological processes involving TEs. Furthermore, it allows the integration of background knowledge.

INTRODUCTION

Currently, the annotation of TEs involves a fair amount of manual labor. Automated methods exist that screen DNA for candidate TEs, but human annotators take over from there. In this work, we explore how inductive logic programming (ILP) can be used to improve the screening. The framework we propose uses existing methods to create a logic-based representation for each sequence, and then applies an ILP model. In this work, we focus on predicting LTR retrotransposons, a particular type of TEs that is characterized by having long terminal repeats (LTRs) at the boundaries.

METHODS

We propose the following three-step framework [1].

1. The genome is screened for potential LTR retrotransposons. To that aim, we use the tool LTR Finder [2], which scans a DNA sequence to search for matching string pairs (the LTRs), and then filters the list by checking user defined length restrictions. Each remaining candidate, i.e., the region bounded by the LTR pairs, receives a score, depending on how many of a predefined set of structural elements are found in there. The output of this first step is a list of candidate LTR retrotransposons, to be further filtered.
2. Every candidate TE sequence, obtained in the previous step, is screened for the occurrence of protein domains that are known to occur in LTR retrotransposons. Domains are recognized using a profile hidden Markov model (HMM) trained on a multiple sequence alignment corresponding to that subdomain.
3. Each candidate sequence is represented in a first order logic format, by simply listing all its predicted protein domains, and the location in the sequence where that domain was found (see Figure 1). For a given sequence, this representation is fed into an ILP model, together with biological background knowledge. The model predicts for every LTR retrotransposon superfamily the probability that the sequence belongs to that family.

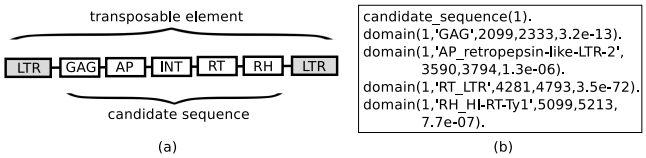


***FIGURE 1**. (a) Typical structure of an LTR retrotransposon, delimited by LTRs and annotated with protein domains. (b) Example of an interpretation, consisting of protein domain predictions. For each domain prediction, we have the candidate ID, the domain, the start and end positions in the sequence, and the e-value for the HMM prediction.*

For the ILP model, which is to be learned from data, the learning process is as follows. For each LTR retrotransposon superfamily, a separate model is learned that maps a sequence, represented as above, to the probability that the sequence belongs to that superfamily. This model is built using the FORF approach (first-order random forests) [3]. The language bias includes the following types of tests that are allowed in the nodes of the trees: (1) the occurrence of a particular protein domain, (2) the occurrence of a particular protein domain before another domain, and (3) the number of occurrences of a particular protein domain. As domains may have subtypes, we give the hierarchical "is a subtype" relationship as background knowledge.

RESULTS

Preliminary results based on precision-recall analysis show a significant improvement over state-of-the-art techniques.

REFERENCES

1. E. Costa, L. Schietgat, R. Cerri, C. Vens, C. Fischer, C. Carareto, J. Ramon, and H. Blockeel, Annotating transposable elements in the genome using relational decision tree ensembles, 23rd International Conference on Inductive Logic Programming (2013)
2. Xu, Z., Wang, H.: LTR FINDER: an efficient tool for the prediction of full-length LTR retrotransposons. Nucleic Acids Research 35(suppl 2) (2007)
3. Van Assche, A., Vens, C., Blockeel, H., Dzeroski, S.: First order random forests: Learning relational classifiers with complex aggregates. Machine Learning 64(1-3) (2006)

SMALL DECISION MODELS: CAPITALIZING ON FEATURE SELECTION

Lennart Backus[1,2,*], *Jos Boekhorst*[1,4], *Sacha van Hijum*[1,2,3].

Centre for Molecular and Biomolecular Informatics, Radboud University Nijmegen Medical Centre [1], P.O. Box 9101, Nijmegen, the Netherlands; TI Food and Nutrition [2], P.O. Box 557, 6700 AN Wageningen, The Netherlands; Netherlands Bioinformatics Centre [3], P.O. Box 9101, 6500 HB Nijmegen, the Netherlands; NIZO food research [4], P.O. Box 20, 6710 BA Ede, The Netherlands

Within the field of bioinformatics samples from ~omics data can often be categorized in groups based on e.g. sample origin or observed phenotype. Often the goal is to identify features (e.g. genes) that differ between the sample groups. These ranked lists of features (based on importance or significance) can be generated by various methods such as Random Forest (RF), Support Vector Machines (SVM), or even simple statistic tests such as a students' *t*-test. The challenge lies in interpreting these lists and understanding how (combinations of) features exactly associate to the sample groups. This is currently a time consuming process but can lead to significant biological insights and aid in determining suitable follow-up experiments. We propose a method and a tool that can create Simple Decision Models (SDMs) that describe the important features in much higher detail; these models are a major improvement over plain feature lists. We demonstrate this by creating a SDM for a metagenomic dataset thereby uncovering detailed feature information.

INTRODUCTION

In bioinformatics often the measurements of samples of large ~omics datasets are used to train a classification model based on a given sample grouping. Classification models can be used to predict the class of new samples based on a trained model. These models can be used to determine lists of important features that allow separating the sample groups. Interpreting these lists of features is the first step towards understanding the underlying biological mechanism and currently is a tedious task.

Small Decision Models (SDMs), decision trees consisting of a small number (< 8) of features, allow for an intuitive way of visualizing how a feature and its cut-off relate to the classification groups. We created the RFScout tool to construct SDMs from almost any type of dataset based on Random Forest (RF)[1]. It enables the user (biologist or bioinformatician) to create SDMs manually, guided by information gathered from RF-trained models.

METHODS

RFscout guides the user in creating SDMs based on a given dataset. The tool is based on RF and its useful features[2] such as capturing potential interactions between features. This allows the user to create SDMs using (i) *Dynamic feature importance information*: feature importance for a (sub-) group of samples. (ii) *Cut-off value suggestion*: at what threshold will a feature be specific for a classification group? (iii) *Combinatorial features*: find combinations of features that only together become highly predictive. (iv) *Sample stratification detection*: finding smaller sub-classification groups of samples.

Here the methodology is applied on a metagenomics dataset querying the microbiome composition on different locations of the body of nine human subjects[3]. The samples are classified based on the gender of the host. Samples from two locations are included (forehead and tongue). The goal is to see which bacteria (OTUs) are specific for male or female hosts and uncover "hidden" sub-groups of samples for both the locations (forehead and tongue).

RESULTS & DISCUSSION

Based on the classes male and female, we constructed an SDM with RFscout consisting of three bacteria (Figure 1). It provides detailed information on: (i) cut-off values, which are for this dataset generally low (< 0.1%) signifying presence / absence of specific bacteria. (ii) Relation of (combinatorial) features to classification groups, for example if OTU_120 and OTU_1449 are both present the sample is classified as a male forehead. (iii) Stratification, only two classification groups (male and female) are used for classification yet we find four sample groups in the decision model which correlate with the location in both male and female.

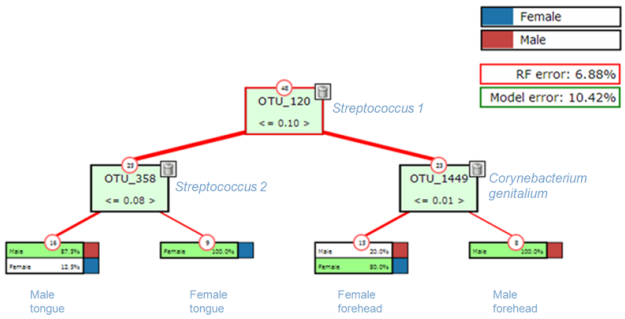


***FIGURE 1**. SDM for classifying the gender of, and location on, the host of metagenomics samples.*

Presenting these complex relations in an intuitive model allows the user to find and describe the important features in their dataset in relation to samples on a much more detailed level than plain lists. This allows for more informed decision making for experimental validation of leads.

REFERENCES

1. Breiman L. Machine Learning **45**, 5-32 (2001).
2.Touw WG et al. Briefings in bioinformatics **14**, 315-326 (2013).
3.Costello EK et al. Science **326**, 1694-7 (2009).

PREDICTING TRYPTIC CLEAVAGE FROM PROTEOMICS DATA USING DECISION TREE ENSEMBLES

Thomas Fannes[1,], Elien Vandemarliere[23], Leander Schietgat[1], Sven Degroeve[23], Kurt De Grave[1], Lennart Martens[23] & Jan Ramon[1].*

Dept. of Computer Science[1], KU Leuven; Dept. of Medical Protein Research[2], VIB; Dept. of Biochemistry[3], Ghent University. [*]thomas.fannes@cs.kuleuven.be

Trypsin is the workhorse protease in mass spectrometry-based proteomics experiments and is used to digest proteins into more readily analyzable peptides. To identify these peptides after mass spectrometric analysis, the actual digestion has to be mimicked as faithfully as possible *in silico*. We introduce CP-DT (Cleavage Prediction with Decision Trees), an algorithm based on a decision tree ensemble that was learned on publicly available peptide identification data from the PRIDE repository. CP-DT is able to accurately predict tryptic cleavage: tests on three independent data sets show that CP-DT significantly outperforms the Keil rules that are currently used to predict tryptic cleavage. Moreover, the trees generated by CP-DT can make predictions efficiently and are interpretable by domain experts.

INTRODUCTION

Trypsin is the most used enzyme in proteomics experiments to convert proteins into peptides as it has a high substrate specificity: it cuts exclusively after arginine and lysine residues. A typical problem is to identify an unknown protein: the protein is cleaved with trypsin and after mass spectrometry, the resulting spectra are compared to theoretical spectra to allow for an identification of the unknown peptides and thus of the unknown protein. The size of the search space is dependent on the number of possible peptides which is quadratic in the number of possible cleavage positions. Accurately predicting cleavage or miscleavage thus reduces the search space.

METHODS

In our work[1] we use machine learning techniques to learn a model capable of predicting trypsin cleavage based on the primary structure of a protein and a possible cut position in the sequence. We allow a number of tests on the amino acids type and/or their properties within a window around the possible cut position, e.g., "Is there an amino acid with neutral charge two positions after the cut position?" or "Is there a proline within distance one of the cut position?" As tryptic cleavage is highly localized, tests are restricted to a window of width 6 centered around the possible cleavage position. We learn a decision tree ensemble, in particular a set of 100 decision trees where each new node in a tree is selected from a random subset of the available tests rather than all. The prediction of the ensemble is generated by aggregating the predicted values of all trees.

The forest was learned on a homogeneous dataset retrieved from PRIDE by selecting all 681,193 examples containing trypsin cleavage information. The predictive performance was estimated using cross-validation on the training set and on three independent test sets: the iPRG-dataset (9,694 examples), the CPTAC-dataset (23,842 examples) and the MS LIMS-dataset (26,079 examples). We compare our model with respect to an existing set of rules, the so called "Keil rules". We also evaluated models specific for a single species, Arg binding sites, or Lys binding sites.

RESULTS & DISCUSSION

Evaluated on the PRIDE dataset, CP-DT attains an AUROC of 96%, an improvement of 28% with respect to the Keil rules. On the three independent datasets, our method achieves AUROC scores of 83% to 90%, significantly outperforming the Keil rules with an average improvement in AUROC of 18%. We therefore conclude that our trypsin cleavage predictor favorably compares to the state-of-the-art model.

No significant differences in enzyme specificity across species were found. Models learned on the Lys subset have a similar quality in predicting Lys and Arg, but models learned on Arg perform significantly worse on the Lys subsets.

URL

CP-DT is available at http://dtai.cs.kuleuven.be/trypsin

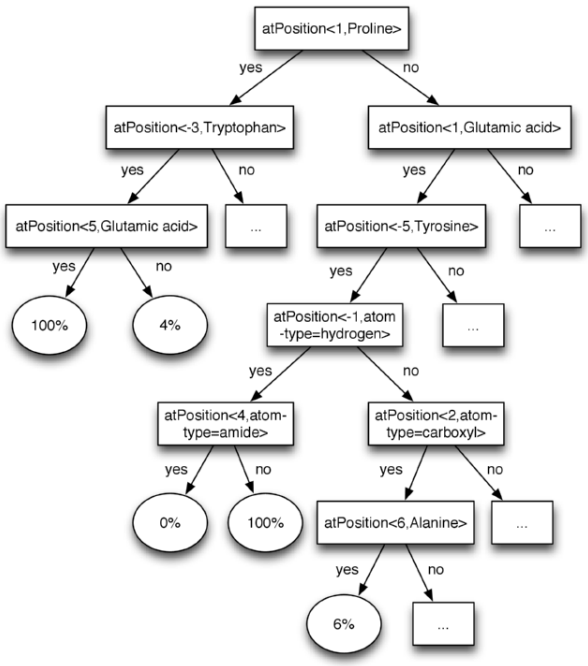


***FIGURE 1**. Example of a decision tree predicting the probability that a protein will be cleaved.*

REFERENCES

1. Fannes T *et al. J Proteome Research* **12**, 2253-2259 (2013).

PROMOTING A FUNCTIONAL AND COMPARATIVE UNDERSTANDING OF THE CONIFER GENOME-IMPLEMENTING APPLIED ASPECTS FOR MORE PRODUCTIVE AND ADAPTED FORESTS

Lieven Sterck[1,2,], Zhen Li[1,2], Yves Van de Peer[1,2] & ProCoGen EU consortium*
[1] *Department of Plant Systems Biology, VIB, Technologiepark 927, 9052 Ghent, Belgium.*
[2] *Department of Plant Biotechnology and Bioinformatics, Ghent University, Technologiepark 927, 9052 Ghent, Belgium.*
[*] *lieven.sterck@psb.vib-ugent.be*

Conifers are the target of major tree breeding efforts worldwide and in the midst of a climatic change scenario the genetics of adaptive response in conifers becomes essential to ensure a sustainable management of genetic resources and effective breeding.

INTRODUCTION

Advances in molecular technologies, such as next-generation DNA sequencing, could have an enormous impact on the rate of progress and achievements made by tree breeding programmes. These new technologies might be used not only to improve our understanding of fundamental conifer biology, but also to address practical problems for the forest industry as well as problems related to the adaptation and management of conifer forests. In this context, ProCoGen will address genome sequencing of two keystone European conifer species. Genome re-sequencing approaches will be used to obtain two reference pine genomes. Comparative genomics and genetic diversity will be closely integrated and linked to targeted functional genomics investigations to identify genes and gene networks that efficiently help to develop or enhance applications related to forest productivity, forest stewardship in response to environmental change or conservation efforts. The development of high-throughput genotyping tools will produce an array of pre-breeding tools to be implemented in forest tree breeding programmes. ProCoGen will also develop comparative studies based on orthologous sequences, genes and markers, which will allow guiding re-sequencing initiatives and exploiting the research accumulated on each of the species under consideration to accelerate the use of genomic tools in diverse species. ProCoGen will integrate fragmented activities developed by European research groups involved in several on-going international conifer genome initiatives and contribute to strengthening international collaboration with North American initiatives (US and Canada).

OBJECTIVES

To develop integrative and multidisciplinary genomic research in conifers, using high-throughput platforms for sequencing, genotyping and functional analysis and to unravel genome organization to identify genes and gene networks controlling important ecological and economic traits, such as those related to control and reduction of climatic change impact in relation to growth, drought and cold stress.

***FIGURE 1**. The four conifer species being studied within the ProCoGen project.*

NMR_REDO: LARGE-SCALE RECALCULATION OF NMR STRUCTURES IN THE CLOUD

Touw W.G.[1,*], *Doreleijers J.F.*[1], *Vuister G.W*[2]. *& Vriend G.*[1,*]
Centre for Molecular and Biomolecular Informatics (CMBI), NCMLS, RadboudUMC, NL[1], *Department of Biochemistry, University of Leicester, UK*[2]. **w.touw@umcn.nl, g.vriend@umcn.nl*

The determination of solution structures of proteins by NMR is an elaborate process with many imperfect experimental and computational steps that often are based on largely empirically determined procedures. Over time, these procedures have become more advanced. By applying today's improved technology to the original data, better structures can be calculated, especially when the original structure was calculated 10 or 20 years ago.

INTRODUCTION

NMR_REDO aims to improve quality of NMR-derived biomolecular structure ensembles in terms of fit with the experimental data and geometric quality by automated state-of-the-art recalculations. Furthermore, the NMR_REDO database may serve as a baseline for methodological improvements. Finally, NMR_REDO aims to generate representative ensembles that properly reflect data uncertainty and dynamic processes.

METHODS

The NMR_REDO pipeline runs fully automatically in the cloud and consists of three main steps for each structure.

- A CING[1] project with all necessary data is retrieved from the NRG-CING[2] database and prepared for
- Simulated annealing with Xplor-NIH and water refinement[3,4].
- The final ensemble is validated and stored in the NMR_REDO database.

RESULTS & DISCUSSION

Almost 3400 NMR structures have been recalculated with our current protocol, consuming 100 000 hours of CPU time. We estimate we will be able to redo about 5000 structures when our protocol can deal with orientational restraints, ligands, multimers and heterogeneity of data formats. NMR_REDO ensembles on average show a better fit to the experimental data and to independent validation criteria than the original NMR structures. Furthermore, NMR_REDO structures are more similar to crystal structures of the same protein.

***FIGURE 1**. In the early days of NMR, occasionally something went wrong.*

REFERENCES

1. Doreleijers JF *et al. J Biol NMR* **54**, 267-83 (2012).
2. Doreleijers JF *et al. Nucleic Acids Res.* **40**, D519–24 (2012).
3. Nabuurs SB *et al. Proteins* **55**, 483-6 (2004).
4. Nederveen AJ *et al. Proteins* **59**, 662-72 (2005).

COMPARATIVE MOTIF DISCOVERY IN THE CLOUD

Dieter De Witte[*,1], *Jan Van de Velde*[2,3], *Michiel Van Bel*[2,3], *Pieter Audenaert*[1], *Piet Demeester*[1], *Bart Dhoedt*[1], *Klaas Vandepoele*[2,3] *and Jan Fostier*[1]

Dept. of Information Technology (INTEC), Ghent University[1], *Dept. of Plant Systems Biology, VIB, Ghent*[2], *Dept. of Plant Biotechnology and Bioinformatics, Ghent University*[3]. [*]*dieter.dewitte@intec.ugent.be*

We present a novel method for the computational discovery of cis-regulatory elements ('motifs') in genomic sequences based on phylogenetic footprinting. Word-based, exhaustive approaches are among the best performing methods; however, they pose significant computational challenges as the number of candidate motifs to evaluate is very high. We describe a parallel, distributed-memory algorithm for the *de novo* comparative motif discovery that has been implemented in Hadoop's MapReduce framework in order to make efficient use of cloud computing infrastructure. It is used in a comparative study of four Monocot plant species and is able to statistically evaluate the conservation of billions of candidate motifs in less than 34 hours using 20 node instances on the Amazon EC2.

INTRODUCTION

Many computational methods exist for the discovery of cis-regulatory elements, see e.g. [1] for an overview. As sequence information is available for an increasing number of organisms, methods based on phylogenetic footprinting are becoming increasingly attractive. In this contribution, we present such method, called BLSSpeller, with three important and unique features. First, it relies on a word-based methodology in which *all* words that occur in any of the sequences are exhaustively screened for conservation. In contrast to statistical methods, the method yields complete and optimal results. Second, the algorithm does not rely on pre-generated multiple sequence alignments (MSA). This ensures the method is suitable even for diverged species for which the generation of a MSA is difficult. Third, in order to deal with the very high runtimes and memory requirements associated with exhaustive, word-based methodologies, the algorithm is implemented in the MapReduce framework, in order to take advantage of cloud infrastructure such as the Amazon EC2.

METHODS

Orthologous and paralogous genes from related species are grouped into gene families. The promoter sequences are extracted and serve as input for the algorithm. For each gene family individually, all words that are conserved within that family are exhaustively enumerated using Sagot's algorithm [2]. Conservation is scored using the Branch Length Score (BLS) [3] which takes the relative evolutionary distance between the organisms into account. Words are enumerated in the IUPAC alphabet with the exclusion of three-fold degenerate characters (B, D, H, V). For each word, the number of gene families in which the word is conserved with a BLS higher than a pre-specified threshold is counted. This number is compared to a background model, i.e., the expected number of gene families in which permutations of that word (i.e., same length, base pair composition and degeneracy) is conserved. For each word, a confidence score C is established, only motifs with a confidence $C > 0.9$ are retained [3].

In order to deal with the very high number of words, the algorithm was implemented using the MapReduce framework. During the map phase, different gene families are attributed to different map tasks and each map task enumerates all conserved words. As the number of words produced by a map task is typically too high to store in memory, the words are written to local disc. In the reduce phase, the conservation score C is established for each word. In between the map and reduce phase, words are sorted on disc in a distributed fashion.

RESULTS & DISCUSSION

The input consists of four related Monocot species: *Zea mays*, *Sorghum bicolor*, *Brachypodium distachyon* and *Oryza sativa*. Using the integrative orthology method of Plaza 2.5 [4], 17724 gene families were constructed. For each gene, the 2kbp promoter was extracted. The software was run on the Amazon EC2 cloud on 20 nodes (type m1.xlarge). The total runtime was 33 hours and 28 minutes. In total, over 2.4 trillion candidate motifs with a length between 6 to 12 basepairs and a maximum of three degenerate characters were emitted by the mappers. This corresponds to 3.65 TByte of data that was sorted on disc in between the map and reduce phase. Over 620 million words with a confidence score $C > 0.9$ were retained by the reduce tasks. This corresponds to roughly 4% of the total number of unique words examined. The identified motif instances show a significant enrichment towards (experimentally characterized) open chromatin regions in rice seedling (12,59 fold enrichment, $p < 0.001$), revealing their biological relevance. Additionally, the method correctly predicts a set of experimentally determined Knotted1 gene targets that were obtained using ChIP-Seq combined with transcript profiling in *Zea mays* [5]. Future work consists of the development of post-processing tools by which the output of the discovery algorithm can be queried in order to identify specific regulatory interactions.

REFERENCES

1. Das M.K. & Dai H.-K. *BMC bioinformatics* **8** Suppl. 7, S21 (2007).
2. Marsan L. and Sagot M. F. Journal of computational biology 7, 325-362 (2000).
3. Kheradpour P. *et al. Genome research* **17**(12), 1919-1931 (2007).
4. Van Bel *et al. Plant Physiology*, 111.189514 (2011).
5. Bolduc N. *et al. Genes Dev.* **26**(15), 1685-1690, (2012).

LARGE SCALE ANCESTRAL ROUTE RECONSTRUCTION OF THE LUXEMBOURGIAN HIV COHORT IN INTERNATIONAL CONTEXT

Daniel Struck[1,*], *Danielle Perez Bercoff*[1], *Carole Seguin-Devaux*[1], *Jean-Claude Schmit*[1].
Laboratory of Retrovirology, CRP-Santé, Luxembourg[1]. [*]*daniel.struck@crp-sante.lu*

Large scale phylogenetic analyses using powerful bioinformatic tools can provide new insights in the dynamics of epidemic infectious diseases. Here we investigated the dynamics of HIV infection in Luxembourg by determining its geographical origin, route of transmission, of entry into the country and expansion of foci within Luxembourg among different risk groups.

MATERIALS & METHODS

A codon corrected multiple alignment was generated from 31430 HIV-1 prot-RT sequences from the LANL database and 601 sequences from the Luxembourg (lux.) cohort (minimum length >1245 bp). The lux. cohort is composed of 159 female and 442 male patients. The contamination routes were 277 heterosexual (46.09%), 241 MSM (40.1%), 62 IVDU (10.32%), 11 unknown (1.83%) and 10 other (1.66%). Sequences from the lux cohort were labeled with the contamination route. Sequences from the LANL database were labeled "external".

A first tree including all Luxembourg and LANL sequences was generated with FastTree. The lux. cluster distribution was analyzed with a custom script using the biopython library.

In a second analysis, 100 bootstrapped phylogenetic trees were inferred from the initial alignment with FastTree. The ancestral states, i.e. the contamination routes, were reconstructed with Mesquite. The trees mapping the ancestral routes were exported and analyzed with in-house scripts. Transitions are inter-group infections within the lux. cohort.

RESULTS

The cluster analysis showed that the majority of infections are single point introductions (343 out of 601, 57.07%) that do not spread further within the lux. cohort. Clusters of more than 5 patients were: 1) one cluster of 37 patients infected with CRF42_BF through heterosexual and homosexual contact, 2) 8 clusters of mainly men, 7 including only MSMs and 1 including MSM and heterosexuals, 3) one cluster of IVDUs (9 patients) and 2 small heterosexual clusters. In 39 cases (13.64%) the virus spreaded within likely couples (cluster size of 2) (23 male/female, 16 MSM).

Reconstruction of the ancestral route showed that the HIV epidemic in Luxembourg is impacted by substantial external inputs: 362 (average: 100 boostrap values, standard deviation (std) 4.61, 60.17%) new infections were derived from an ‘external’ ancestor vs. 239 (average: std 4.61, 39.83%) from internal routes. Most infections with an 'external' ancestor occurred through heterosexual contact, immediately followed by homosexual contact and then by IVDU. When the analysis was restricted to patients born in Luxembourg only, the distribution shifted to 97.29 (std: 2.81, 52.59%) external vs. 87.71 internal routes (std: 2.81, 47.41%). The average number of transitions between 'internal' transmissions was 9.92% for the whole cohort and were mainly from MSM to heterosexual and from heterosexuals to MSMs.

CONCLUSIONS

Large scale ancestral reconstruction of the contamination routes using viral sequences showed that the lux. cohort is mainly driven by single-point external introductions through heterosexual and homosexual contacts leading to a dead-end. Only few large clusters exist in the lux. cohort, due mainly to homosexual transmissions. This study shows that fine-grained analyses of infectious diseases are possible and can be exploited to target prevention strategies.

SPECIES INTERACTIONS IN THE WORLD'S OCEANS

Gipsi Lima-Mendez *[1,2], Karoline Faust [1,2], Nicolas Henry[3,4,5], Sebastien Colin[3,4,5], Stephane Audic[3,4,5], Shinichi Sunagawa [6], Silvia G. Acinas[7], Hiroyuki Ogata[8], Patrick Wincker[9], Chris Bowler [10], Colomban de Vargas [3,4,5], Eric Karsenti [6], Jeroen Raes[1,2] and Tara-Oceans Consortium.

*Dept. of Structural Biology, VIB[1], Dept. of Applied Biological Sciences (DBIT), Vrije Universiteit Brussel[2], CNRS[3], Université Pierre et Marie Curie (Paris 06), UMR 7144 [4], Station Biologique de Roscoff 5, European Molecular Biology Laboratory6, Department of Marine Biology and Oceanography, Institute of Marine Science (ICM), CSIC[7], Education Academy of Computational Life Sciences, Tokyo Institute of Technology [8], CEA, Institut de Génomique, Genoscope [9], Institut de Biologie de l'Ecole Normale Supérieure (IBENS) UMR8197 3, *gipsi.lima@vub.ac.be*

The structure and dynamics of the oceans result from the interplay between the inhabitant organisms and the physical component of the environment. Using a dataset collected from 47 diverse sites during the Tara Oceans expedition we infer networks based on the taxon abundance profiles. Edges represent pair-wise relationships between taxonomic units that tend to co-occur across different samples. The analysis of these pair-wise links given the environmental conditions across the samples revealed that the associations were either not affected or enhanced when corrected by the environmental factors.

INTRODUCTION

In the oceans, the basis of the food web is provided by the plankton, which includes all drifting organisms, e.g. viruses, bacteria, archaea, protists and even small metazoan like jellyfish and fish larvae. The Tara Oceans expedition collected, between 2009 and 2012, planktonic organisms together with physico-chemical contextual data across the world oceans. Plankton were collected at three different depths: near the surface, at the depth of maximum chlorophyll *a* fluorescence (Deep Chlorophyll Maximum, DCM) and in the mesopelagic layer. The water samples were size-fractionated to cover organisms of sizes spanning from 0.2-2000 µm.

The occurrence profiles of species or operationally taxonomic units (OTU) across the samples provide insights into their ecological relationships. We aim at building a global, cross-kingdom species co-occcurence network, and disentangle interactions that are due to abiotic factors (niche, environment) and those that might represent true biotic interactions (symbiotic, competitive, predation).

METHODS

We used the OTU abundances derived from V9 18S rRNA pyrotags and 16S RNA metagenomic reads from 47 sample sites to infer networks using the software CoNet [1].

To disentangle the effects of biotic and abiotic factors we calculated the interaction information [2] for each edge and environmental variable combination:

$$I(X,Y,Z) = I(X,Y \mid Z) - I(X,Y)$$

where X and Y are OTU abundance vectors and Z is the environmental variable vector and I is the mutual information.

A positive interaction information value indicates that the dependency between taxon X and Y is enhanced by the environmental factor Z, a negative value indicates redundancy, i.e. the dependency between taxon X and Y can be totally explained by the factor Z and a value of zero denotes that Z does not modify the interaction between X and Y.

The environmental parameter tested were latitude, longitude, salinity, oxygen, nitrate, nitrite, chlorophyll a, dissolved organic matter, particle specific measures, PO4, and SI.

RESULTS & DISCUSSION

The network built from the 18S derived OTUs consists of 1042 nodes interconnected by 2728 edges: 1043 copresences and 1685 exclusions. Most edges (2623, 96%) are fraction and depth specific. The prokaryotic network contains 3492 edges, of which only 610 are common to the surface and the DCM layers. This structure clearly reflects the stratification of the ocean communities by depth.

The environment enhanced the dependency of 656 out of the 1042 edges in the eukaryotic network (positive interaction information). In the prokaryotic network we found 1930 edges to have positive interaction information with at least one environmental parameter. Only 4 values were negative; salinity could explain one co-occurring pair in the eukaryotic network and oxygen could explain 3 of co-occurring prokaryotic genera (Figure 1) and nitrite could explain the exclusion between 2 prokaryotic genera.

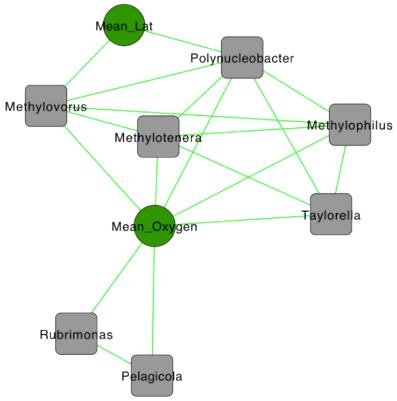


***FIGURE 1**. Positively correlated genera explained by a common correlation to oxygen.*

REFERENCES

1. Faust, K. et al. *PLoS Comput Biol* **8**, e1002606, (2012).
2. Jakulin, A. & Bratko, I. in *Twenty-First International Conference on Machine Learning* (ICML-2004). eds R. Greiner & D. Schuurmans) 409-416 (2004).

PARTICIPANTS LIST

LAST NAME	FIRST NAME	EMAIL	AFFILIATION
Adriaens	Michiel	m.e.adriaens@amc.uva.nl	Academish Medisch Centrum, Amsterdam
Aerts	Stein	stein.aerts@med.kuleuven.be	Katholieke Universiteit Leuven
Albert	Jaroslav	jaroslav.albert@ulb.ac.be	Université Libre de Bruxelles
Ardeshirdavani	Amin	amin.ardeshirdavani@esat.kuleuven.be	Katholieke Universiteit Leuven
Arslan	Ahmed	ahmed.arslan@student.kuleuven.be	Katholieke Universiteit Leuven
Azuaje	Francisco	francisco.azuaje@crp-sante.lu	CRP-Santé, Luxembourg
Backus	Lennart	lennart.backus@radboudumc.nl	Radboud Universitair Medisch Centrum Nijmegen
Basmagi	Said	basmagi.s@hsleiden.nl	Hogeschool Leiden
Baurain	Denis	denis.baurain@ulg.ac.be	Université de Liège
Bayjanov	Jumamurat	j.bayjanov@umcn.nl	Radboud Universitair Medisch Centrum Nijmegen
Beaume	Nicolas	nicolas.beaume@uni.lu	Université du Luxembourg
Belau	Kamil	acidcave@o2.pl	University of Gdansk, Poland
Ben Taieb	Souhaib	bsouhaib@gmail.com	Université Libre de Bruxelles
Berntsen	Karen	k.berntsen@student.science.ru.nl	Radboud Universiteit
Bittremieux	Wout	wout.bittremieux@uantwerpen.be	Universiteit Antwerpen
Bizet	Martin	mbizet@ulb.ac.be	Université Libre de Bruxelles
Boender-van Dijk	Leonie	leonie.boender-van-dijk@dsm.com	DSM Biotechnology Center
Boes	Olivier	olivboes@ulb.ac.be	Université Libre de Bruxelles
Borghei	Mojgansadat	mborghei@ulb.ac.be	Université Libre de Bruxelles
Bottu	Guy	gbottu@vub.ac.be	Vrije Universiteit Brussel, VIB
Branders	Vincent	vbranders@gmail.com	Université Libre de Bruxelles
Brysbaert	Guillaume	guillaume.brysbaert@iri.univ-lille1.fr	Université de Lille 1
Cannoodt	Robrecht	robrecht.cannoodt@ugent.be	Universiteit Gent
Carcillo	Fabrizio	fabrizio.carcillo@ulb.ac.be	Interuniversity Institute of Bioinformatics in Brussels
Cilia	Elisa	ecilia@ulb.ac.be	Université libre de Bruxelles
Croes	Didier	didierc@croes-pasture.be	Vrije Universiteit Brussel, UZ Brussel
Cuypers	Bart	bart.cuypers@uantwerpen.be	Universiteit Antwerpen
D'hoe	Kevin	kevinjex@gmail.com	Katholieke Universiteit Leuven, VUB, VIB
Dalkas	Giorgos	gdalkas@ulb.ac.be	Université Libre de Bruxelles
Dal Pozzolo	Andrea	adalpozz@ulb.ac.be	Université Libre de Bruxelles
Daneels	Dorien	dorien.daneels@uzbrussel.be	Vrije Universiteit Brussel, UZ Brussel
Datema	Erwin	sa@keygene.com	Keygene NV
Davie	Kristofer	kristofer.davie@gmail.com	Katholieke Universiteit Leuven
de Been	Mark	m.debeen-2@umcutrecht.nl	Universitair Medisch Centrum Utrecht
De Grave	Kurt	kurt.degrave@cs.kuleuven.be	Katholieke Universiteit Leuven
de Jonge	Ronnie	ronnie.dejonge@psb.ugent.be	Universiteit Gent, VIB
De Laet	Marie	madelaet@ulb.ac.be	Université Libre de Bruxelles
De Maeyer	Dries	dries.demaeyer@biw.kuleuven.be	Katholieke Universiteit Leuven
De Meyer	Tim	tim.demeyer@ugent.be	Universiteit Gent
De Schrijver	Joachim	joachim.deschrijver@multiplicom.com	Multiplicom NV
De Smet	Riet	riet.desmet@psb.vib-ugent.be	Universiteit Gent, VIB
De Smet	Matthias	matthias.de.smet@me.com	Universiteit Gent

De Witte	Dieter	drdwitte@gmail.com	Universiteit Gent
Defrance	Matthieu	matthieu.dc.defrance@ulb.ac.be	Université Libre de Bruxelles
Degroeve	Sven	sven.degroeve@ugent.be	Universiteit Gent, VIB
Dehouck	Yves	ydehouck@ulb.ac.be	Université Libre de Bruxelles
Delcourt	Thomas	thomas.delcourt@student.uclouvain.be	Université Libre de Bruxelles
Di Franco	Arnaud	arnaud.difranco@gmail.com	Université de Liège
Dutilh	Bas	bedutilh@gmail.com	Radboud Universitair Medisch Centrum Nijmegen
Eijssen	Lars	l.eijssen@maastrichtuniversity.nl	Maastricht University
El Aalamat	Yousef	yousef.elaalamat@esat.kuleuven:be	Katholieke Universiteit Leuven
ElShal	Sarah	sarah.elshal@esat.kuleuven.be	Katholieke Universiteit Leuven
Escaliere	Bertrand	bertrand.escaliere@gmail.com	Interuniversity Institute of Bioinformatics in Brussels
Essaghir	Ahmed	ahmed.essaghir@uclouvain.be	Université Catholique de Louvain
Estievenart	Quentin	qestieve@ulb.ac.be	Université Libre de Bruxelles
Evelo	Chris	chris.evelo@maastrichtuniversity.nl	Maastricht University
Faust	Karoline	karoline.faust@vib-vub.be	Vrije Universiteit Brussel, VIB
Feenstra	Anton	k.a.feenstra@vu.nl	Vrije Universiteit Amsterdam
Fiers	Mark	mark.fiers@cme.vib-kuleuven.be	Katholieke Universiteit Leuven, VIB
Fostier	Jan	jan.fostier@intec.ugent.be	Universiteit Gent
Gaigneaux	Anthoula	anthoula.gaigneaux@lbmcc.lu	LMBCC Luxembourg
Galhardo	Mafalda	mafalda.galhardo@uni.lu	Université du Luxembourg
Gazzo	Andrea	andrea.gazzo.86@gmail.com	Interuniversity Institute of Bioinformatics in Brussels
Geens	Wouter	wouter.geens@student.kuleuven.be	Katholieke Universiteit Leuven
Georgatos	Fotis	fotis.georgatos@uni.lu	Université du Luxembourg
Gilis	Dimitri	dimitri.gilis@ulb.ac.be	Université Libre de Bruxelles
Godard	Patrice	patrice.godard@thomsonreuters.com	Thomson Reuters
Gonnelli	Giulia	giulia.gonnelli@ugent.be	Universiteit Gent, VIB
Gonze	Didier	dgonze@ulb.ac.be	Université Libre de Bruxelles
Guharoy	Mainak	mainak.guharoy@gmail.com	Vrije Universiteit Brussel, VIB
Haseloff	Jim	jh295@cam.ac.uk	University of Cambridge
Hayes	Matthew	mhayes2@mcneese.edu	McNeese State University
Hendrickx	Diana	d.hendrickx@maastrichtuniversity.nl	Maastricht University
Hermans	Susanne	susanne.hermans@gmail.com	Radboud Universitair Medisch Centrum Nijmegen
Housset	Nicolas	nicolas.housset@ugent.be	Universiteit Gent, VIB
Illeghems	Koen	koen.illeghems@vub.ac.be	Vrije Universiteit Brussel
Imrichová	Hana	hana.imrichova@med.kuleuven.be	Katholieke Universiteit Leuven
Jacobs	Jelle	jelle.jacobs@student.kuleuven.be	Katholieke Universiteit Leuven
Janky	Rekin's	rekins.janky@med.kuleuven.be	Katholieke Universiteit Leuven
Kalender Atak	Zeynep	zeynep.kalender@med.kuleuven.be	Katholieke Universiteit Leuven
Kiekens	Raphael	raphael.kiekens@howest.be	Hogeschool West-Vlaanderen
Krammer	Eva-Maria	ekrammer@ulb.ac.be	Université Libre de Bruxelles
Kumar	Ajay	ajay.kumar@uantwerpen.be	Universiteit Antwerpen
Kwasigroch	Jean Marc	jean.marc.kwasigroch@ulb.ac.be	Université Libre de Bruxelles
Laenen	Griet	griet.laenen@esat.kuleuven.be	Katholieke Universiteit Leuven
Laukens	Kris	kris.laukens@uantwerpen.be	Universiteit Antwerpen
Le Van	Thanh	thanh.levan@cs.kuleuven.be	Katholieke Universiteit Leuven
Lenaerts	Tom	tlenaert@ulb.ac.be	Université Libre de Bruxelles
Lenoci	Leonardo	ll79@member.fsf.org	Radboud Universitair Medisch Centrum Nijmegen

Lensink	Marc	marc.lensink@iri.univ-lille1.fr	Université de Lille 1
Lima Mendez	Gipsi	gipsi.lima@vub.ac.be	Vrije Universiteit Brussel, VIB
Lin	Yao-Cheng	yalin@psb.vib-ugent.be	Universiteit Gent, VIB
Lopes	Miguel	miguelaglopes@gmail.com	Université Libre de Bruxelles
Macossay Castillo	Mauricio	mmacossa@vub.ac.be	Vrije Universiteit Brussel
Marée	Raphaël	raphael.maree@ulg.ac.be	Université de Liège
Marien	Koen	marien@histogenex.com	HistoGeneX
Masuzzo	Paola	paola.masuzzo@ugent.be	Universiteit Gent, VIB
Mensaert	Klaas	klaas.mensaert@ugent.be	Universiteit Gent
Menschaert	Gerben	gerben.menschaert@gmail.com	Universiteit Gent
Meyer	Patrick	pmeyer@ulb.ac.be	Université Libre de Bruxelles
Meysman	Pieter	pieter.meysman@uantwerpen.be	Universiteit Antwerpen
Mobegi	Fredrick	fredrickmaati@gmail.com	Radboud Universitair Medisch Centrum Nijmegen
Moreau	Yves	yves.moreau@esat.kuleuven.be	Katholieke Universiteit Leuven
Moschopoulos	Charalampos	charalampos.moschopoulos@esat.kuleuven.be	Katholieke Universiteit Leuven
Munar	Marta	martaamunar@gmail.com	Vrije Universiteit Brussel
Mysara	Mohamed	mohamed.mysara@sckcen.be	Vrije Universiteit Brussel
Nakano	Yuka	okayuka531@gmail.com	Radboud Universitair Medisch Centrum Nijmegen
Naulaerts	Stefan	stefan.naulaerts@uantwerpen.be	Universiteit Antwerpen
Nazarov	Petr	petr.nazarov@gmail.com	CRP-Santé, Luxembourg
Ndah	Elvis	elvis.ndah@ugent.be	Universiteit Gent
Olsen	Catharina	colsen@ulb.ac.be	Université Libre de Bruxelles
Overkleeft	Rick	s1065247@student.hsleiden.nl	Hogeschool Leiden
Pancsa	Rita	rpancsa@vub.ac.be	Vrije Universiteit Brussel, VIB
Pepe	Daniele	daniele.pepe@esat.kuleuven.be	Katholieke Universiteit Leuven
Porretta	Luciano	lporrett@ulb.ac.be	Université Libre de Bruxelles
Pucci	Fabrizio	fapucci@ulb.ac.be	Université Libre de Bruxelles
Raimondi	Daniele	eddiewrc@alice.it	Interuniversity Institute of Bioinformatics in Brussels
Reggiani	Claudio	claudio.reggiani@ulb.ac.be	Université Libre de Bruxelles
Rooman	Marianne	mrooman@ulb.ac.be	Université Libre de Bruxelles
Ruelle	Jean-Louis	jean-louis.ruelle@gsk.com	GlaxoSmithKline
Ruyssinck	Joeri	joeri.ruyssinck@intec.ugent.be	Universiteit Gent, iMinds
Sabaghian	Ehsan	ehsan.sabaghian@psb.vib-ugent.be	Universiteit Gent, VIB
Saeys	Yvan	yvan.saeys@ugent.be	Universiteit Gent
Sakai	Ryo	ryo.sakai@esat.kuleuven.be	Katholieke Universiteit Leuven
Sifrim	Alejandro	alejandro.sifrim@esat.kuleuven.be	Katholieke Universiteit Leuven
Simonis	Nicolas	nsimonis@ulb.ac.be	InSilico
Slabbinck	Bram	brsla@psb.ugent.be	Universiteit Gent, VIB
Sterck	Lieven	lieven.sterck@psb.vib-ugent.be	Universiteit Gent, VIB
Steyaert	Sandra	sandra.steyaert@ugent.be	Universiteit Gent
Stock	Michiel	michiel.stock@ugent.be	Universiteit Gent
Struck	Daniel	daniel.struck@crp-sante.lu	CRP-Santé, Luxembourg
Sucaet	Yves	sucaet@histogenex.com	HistoGeneX
Svetlichnyy	Dmitry	dmitry.svetlichnyy@med.kuleuven.be	Katholieke Universiteit Leuven
Tanyalcin	Ibrahim	itanyalc@vub.ac.be	Vrije Universiteit Brussel
Teheux	Fabian	fteheux@ulb.ac.be	Université Libre de Bruxelles
Touw	Wouter	wouter.touw@radboudumc.nl	Radboud Universitair Medisch Centrum Nijmegen

Tramontano	Anna	anna.tramontano@uniroma1.it	Sapienza University of Rome
Turatsinze	Jean-Valery	jturatsi@ulb.ac.be	Université Libre de Bruxelles
Valencia	Alfonso	valencia@cnio.es	Spanish National Cancer Research Center
Vallesc	Mireia	mireia.vallesc@gmail.com	Vrije Universiteit Brussel
van Beusekom	Bart	bartvanbeusekom@gmail.com	Radboud Universitair Medisch Centrum Nijmegen
van den Ham	Henk-Jan	h.j.vandenham@erasmusmc.nl	Erasmus Medisch Centrum Rotterdam
van Dijk	Aalt-Jan	aaltjan.vandijk@wur.nl	Wageningen University and Research Centre
van Driel	Marc	marc.van.driel@nbic.nl	Netherlands Bioinformatics Centre (NBIC)
van Eyll	Jonathan	jonathan.vaneyll@ucb.com	UCB Pharma s.a.
Van Gassen	Sofie	sofie.vangassen@intec.ugent.be	Universiteit Gent
van Gelder	Celia	celia.van.gelder@nbic.nl	Netherlands Bioinformatics Centre (NBIC)
Van Goey	Jeroen	jeroen_van_goey@applied-maths.com	Applied Maths NV
van Gurp	Thomas	t.vangurp@nioo.knaw.nl	NIOO-KNAW
van Hijum	Sacha	sacha.vanhijum@radboudumc.nl	Radboud Universitair Medisch Centrum Nijmegen, NIZO
Vandermaliere	Elien	elien.vandermarliere@ugent.be	Universiteit Gent, VIB
Vandervelde	Alexandra	alexandra.vandervelde@vub.ac.be	Vrije Universiteit Brussel, VIB
Vandeweyer	Geert	geert.vandeweyer@uantwerpen.be	Universiteit Antwerpen
Varadi	Mihaly	mvaradi@vub.ac.be	Vrije Universiteit Brussel, VIB
Varughese	Jobin K.	jobinv@gmail.com	University of Bergen
Vens	Celine	celine.vens@irc.vib-ugent.be	Universiteit Gent, VIB
Verbeeck	Nico	nico.verbeeck@esat.kuleuven.be	Katholieke Universiteit Leuven
Verbeiren	Toni	toni.verbeiren@data-intuitive.com	Katholieke Universiteit Leuven
Vermeire	Tessa	tessa.vermeire@ugent.be	Universiteit Gent, VIB
Vermeirssen	Vanessa	vanessa.vermeirssen@psb.vib-ugent.be	Universiteit Gent, VIB
Vet	Stefan	snjvet@gmail.com	Interuniversity Institute of Bioinformatics in Brussels
Vlassis	Nikos	nikos.vlassis@gmail.com	Université du Luxembourg
Vranken	Wim	wvranken@vub.ac.be	Vrije Universiteit Brussel
Vriend	Gert	gerrit.vriend@radboudumc.nl	Radboud Universitair Medisch Centrum Nijmegen
Waegeman	Willem	willem.waegeman@ugent.be	Universiteit Gent
Weckx	Stefan	stefan.weckx@vub.ac.be	Vrije Universiteit Brussel
Weiss	David	davidweiss@insilicodb.org	InSilico
Wenric	Stephane	s.wenric@ulg.ac.be	Université de Liège
Widar	Julie	WIDARJ@be.ibm.com	IBM
Winand	Raf	raf.winand@esat.kuleuven.be	Katholieke Universiteit Leuven
Yilmaz	Sule	sule.yilmaz@ugent.be	Universiteit Gent, VIB
Zakeri	Pooya	pooya.zakeri@esat.kuleuven.be	Katholieke Universiteit Leuven
Zisis	Ioannis	izisis@ulb.ac.be	Université Libre de Bruxelles
Zsolyomi	Fruzsina	fzsolyom@vub.ac.be	Vrije Universiteit Brussel