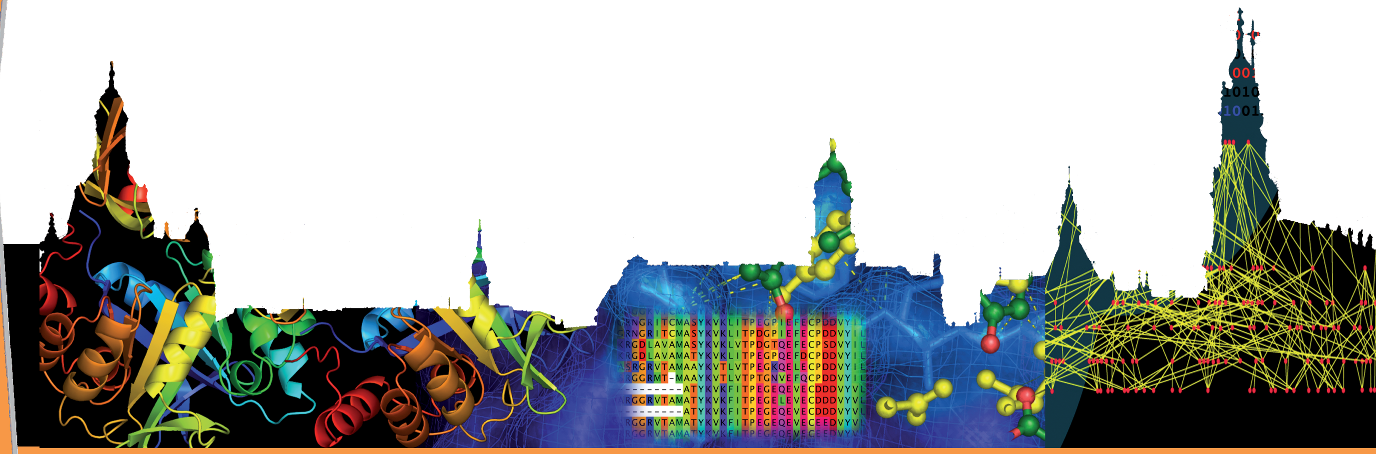
German Conference on Bioinformatics 2008

9. - 12. September 2008

Dresden

Abstracts of Posters and Highlight Talks

Eds: Andreas Beyer and Michael Schroeder



Sponsors of GCB2008

ELSO
European Life Scientist
Organization e.V.

KTS
Klaus Tschira Stiftung
gemeinnützige GmbH

Scientific societies

DECHEMA

GBM

GI

Non-profit societies

Biotechnology Center

CRTD
Center for Regenerative
Therapies Dresden

DEUTSCHES
HYGIENE-MUSEUM
DRESDEN

TECHNISCHE
UNIVERSITÄT
DRESDEN

DIGS-BB

European Commission
Information Society and Media

Commercial sponsors

Cenix
BIOSCIENCE

nAmbition

Scionics
Scionics Computer Innovation GmbH

TRANSINSIGHT

Preface

This volume contains abstracts of posters and highlight papers presented at the German conference on bioinformatics, GCB 2008, held in Dresden, Germany, September 9-12, 2008 at the Deutsches Hygiene Museum Dresden.

GCB is an annual, international conference, which provides a forum for the presentation of current research in bioinformatics and computational biology. It is organized on behalf of the Special Interest Group on Informatics in Biology of the German Society of Computer Science (GI) in cooperation with the German Society of Chemical Technique and Biotechnology (Dechema) and the German Society for Biochemistry and Molecular Biology (GBM) with support of the European Life Science Organization.

GCB 2008 comprises six invited talks by Michael Ashburner, Janusz Bujnicki, David Gilbert, Trey Ideker, Jens Reich and Marino Zerial. The talk by Jens Reich on a person's dignity in the age of the genome chip was co-organised by the Deutsches Hygiene Museum Dresden and GCB. It was held in German and open to the general public. GCB also featured four tutorials by Jens Meiler (Rosetta in computational structural biology), Steffen Möller (expression QTL and their analysis), Johannes Schindelin (image analysis), and Stefan Schuster (metabolic pathway analysis).

Thanks to the programme committee members and reviewers, to the local organizers, and to the sponsors.

Special thanks to Robert Männel for compiling this abstract book.

Dresden, August 2008

Andreas Beyer and Michael Schroeder

Table of Content

Highlight talks Page

Poster

Unexpected complexity at breakpoint junctions in phenotypically normal individuals and mechanisms involved in generating balanced translocations t(1;22)(p36;q13)

Gajecka M*, Gentles AJ, Tsai A, Chitayat D, Mackay KL, Glotzbach CD, Lieber MR, Shaffer LG
School of Molecular Biosciences, Washington State University, Spokane, WA
School of Medicine, Stanford University, Stanford, CA
University of Southern California Comprehensive Cancer Center, University of Southern California, Los Angeles, CA
Prenatal Diagnosis and Medical Genetics, Mount Sinai Hospital, Toronto, Canada
*Corresponding author Email: gajecka@wsu.edu

Abstract:

We investigated breakpoint junctions at the sequence level in phenotypically normal balanced translocation carriers. Eight breakpoint junctions derived from four non-related subjects with apparently balanced translocation t(1;22)(p36;q13) were examined using various molecular biology methods. Next, computational analyses were performed. Additions of nucleotides, deletions, duplications and a triplication identified at the breakpoints demonstrate high complexity at the breakpoint junctions and indicate involvement of multiple mechanisms in the DNA breakage and repair process during translocation formation. Each case presented with different genomic alterations and multiple sequence changes. Sequence motifs and motif densities were evaluated at the breakpoints and junctions of t(1;22). Sequence motifs were found frequently at the breakpoints (oligopurine/oligopyrimidine tracts, vertebrate topoisomerase consensus cleavage sites, DNA polymerase a/b frameshift hotspots and a deletion hotspot consensus sequence, repeats and inverted repeats and other motifs). We calculated the density of each motif in 1kb and 5kb windows to determine the distributions of densities for chromosomes involved in translocations. Except for translin sites in one subject, none of the breakpoint regions show high counts of motifs. Next, to examine whether the sequences near the breakpoints can form cruciform structures, we used UNAFold to produce potential configurations. No putative regions of Z-DNA structure were identified in the breakpoint junctions using Z-Hunt. We postulated that the breakage was owing to random damage (e.g., oxidative damage or ionizing radiation) with no particular predisposition caused by the DNA sequence. We speculate that the sequence complexity found at balanced translocation breakpoints has been underestimated.

Genome Research 2008

Alternative Splicing and Protein Structure Evolution

Fabian Birzele*, Gergely Csaba and Ralf Zimmer

Practical Informatics and Bioinformatics Group, Department of Informatics, Ludwig-Maximilians-University,
Amalienstrasse 17, D-80333 Munich, Germany

*Corresponding author Email: fabian.birzele@bio.ifi.lmu.de

Alternative splicing is thought to be one of the major sources for functional diversity in higher eukaryotes. Interestingly, when mapping splicing events onto protein structures, about half of the events affect structured and even highly conserved regions i.e. are non-trivial on the structure level. This has led to the controversial hypothesis that such splice variants result in nonsense-mediated mRNA decay or non-functional, unstructured proteins, which do not contribute to the functional diversity of an organism. In our study we showed that proteins appear to be much more tolerant to structural deletions, insertions and replacements than previously thought.

We provide examples that splicing events may represent transitions between different folds in the protein sequence–structure space and explain these links by a common genetic mechanism. Taken together, those findings hint to a more prominent role of splicing in protein structure evolution and to a different view of phenotypic plasticity of protein structures.

550–558 Nucleic Acids Research, 2008, Vol. 36, No. 2
doi:10.1093/nar/gkm1054

Published online 30 November 2007

Increasing the mass accuracy of high-resolution LCMS data using background ions – a case study on the LTQ-Orbitrap

Richard A. Scheltema1, Anas Kamleh2, David Wildridge3, Charles Ebikeme3, David G. Watson2, Michael P. Barrett3, Ritsert C. Jansen1 & Rainer Breitling*1

1. Groningen Bioinformatics Centre, Groningen Biomolecular Sciences and Biotechnology Institute, University of Groningen, kerklaan 30, 9751 NN Haren, The Netherlands
2. Strathclyde Institute for Pharmacy and Biomedical Sciences, University of Strathclyde, Glasgow G4 0NR.
3. Institute of Biomedical and Life Sciences, Division of Infection and Immunity, Glasgow Biomedical Research Centre, University of Glasgow, Glasgow G12 8TA.

*Corresponding author Email: r.breitling@rug.nl

Metabolomics is a newcomer to the field of post-genomic technologies. Recent advances in high-accuracy mass spectrometry promise a breakthrough in our ability to analyze complete metabolomes at high sensitivity and reliability [1]. This will allow a system-wide exploration of the structure and function of metabolic networks [2]. This highlight presentation will focus on the novel bioinformatics challenges that we face in high-accuracy metabolomics, ranging from data-preprocessing to metabolic network inference (Scheltema et al., Proteomics, in press). We will illustrate the unprecedented mass accuracy that can be achieved by dedicated algorithms [3], and show how these results can be used in new computational strategies for metabolite identification and the de novo reconstruction of metabolic networks [4, 5].

References

[1] Breitling R, Pitt AR, Barrett MP (2006): Precision mapping of the metabolome. Trends Biotechnol. 24: 543-548.

[2] Breitling R, Vitkup D, Barrett MP (2008): New surveyor tools for charting microbial metabolic maps. Nature Rev. Microbiol. 6: 156-161.

[3] Scheltema RA, Kamleh A, Wildridge D, Ebikeme C, Watson DG, Barrett MP, Jansen RC, Breitling R (2008): Increasing the mass accuracy of high-resolution LC-MS data using background ions – a case study on the LTQ-Orbitrap. Proteomics in press.

[4] Jourdan F, Breitling R, Barrett MP, Gilbert D (2008): MetaNetter: inference and visualization of high-resolution metabolomic networks. Bioinformatics 24: 143-145.

[5] Breitling R, Ritchie S, Goodenowe D, Stewart ML, Barrett MP (2006): Ab initio prediction of metabolic networks using Fourier Transform Mass Spectrometry data. Metabolomics 2: 155-164.

Nature Reviews Microbiology | AOP, published online 19 November 2007;
doi:10.1038/nrmicro1797

Reconstructing protein networks of epithelial differentiation

Niels Grabe[1],[2],*, Thora Pommerencke[1],[2], Thorsten Steinberg[1],[3], Hartmut Dickhaus[1],[2] and Pascal Tomakidi[1],[3]

[1] Hamamatsu Tissue Imaging and Analysis (TIGA) Center, BIOQUANT, University Heidelberg, BQ0010,
Im Neuenheimer Feld 267,
[2] Institute for Medical Biometry and Informatics, University Hospital Heidelberg,
Im Neuenheimer Feld 305 and
[3] Department of Orthodontics and Dentofacial Orthopedics, Dental School University
of Heidelberg, Im Neuenheimer Feld 400, 69120 Heidelberg, Germany

*Corresponding author Email: niels.grabe@bioquant.uni-heidelberg.de

Motivation:
For systems biology of complex stratified epithelia like human epidermis, it will be of particular importance to reconstruct the spatiotemporal gene and protein networks regulating keratinocyte differentiation and homeostasis.

Results:
Inside the epidermis, the differentiation state of individual keratinocytes is correlated with their respective distance from the connective tissue. We here present a novel method to profile this correlation for multiple epithelial protein biomarkers in the form of quantitative spatial profiles. Profiles were computed by applying image processing algorithms to histological sections stained with tri-color indirect immunofluorescence. From the quantitative spatial profiles, reflecting the spatiotemporal changes of protein expression during cellular differentiation, graphs of protein networks were reconstructed.

Conclusion:
Spatiotemporal networks can be used as a means for comparing and interpreting quantitative spatial protein expression profiles obtained from different tissue samples. In combination with automated microscopes, our new method supports the large-scale systems biological analysis of stratified epithelial tissues.

Vol. 23 no. 23 2007, pages 3200–3208
doi:10.1093/bioinformatics/btm504

CELLmicrocosmos 2.1: A Software Approach for the Visualization of three-dimensional PDB Membranes

Björn Sommer*, Sebastian Schneider and Tim Dingersen

Bioinformatics / Medical Informatics Department, Faculty of Technology, University of Bielefeld, D-33594 Bielefeld, Germany

*Corresponding author Email: bjoern@CELLmicrocosmos.org

Background

CELLmicrocosmos is an approach to develop tools for the generation of virtual cell environments. The CELLmicrocosmos 2 project deals with the computational generation of three-dimensional cell membranes.

Biological membranes consist mainly of lipids and proteins. The Protein Data Bank [BWF+00] and the HIC-UP database [KJ98] represent a large number of three-dimensional protein and lipid structures which have been extracted from biological membranes. Other databases contain informations about the membrane-type-specific localization of proteins. There exist various approaches of utilizing these models for the computation of membranes.

Results

Research in many fields of science is dealing with the problem of visualizing, modelling and/or simulating membranes. The theoretical as well as the computational status quo does not allow to generate realistic membranes. Hence, alternatives are created, which are using different developmental environments. Therefore a lot of work has to be invested, before the sophisticated work dealing with algorithms and methods can begin.

We present a software framework, which should allow academics to generate problem-specific membranes: They should be enabled to use simple, short-time as well as complex, time-consuming algorithms featuring a higher grade of realism.

Conclusions

Utilizing Java, Java3D and Jmol [J08], we created a tool which is able to deal with different PDB models, their percental placement and alignment at protein level. A number of algorithms for the lipid placement has been implemented. The most sophisticated one so far is a geometrical-based Monte Carlo algorithm.

References

[BWF+00] H.M. Berman, J. Westbrook, Z. Feng, G. Gilliland, T.N. Bhat, H. Weissig, I.N. Shindyalov and P.E. Bourne. The Protein Data Bank. *Nucleic Acids Research,* 28:235–242. Oxford University Press, Oxford, 2000.

[J08] Jmol: An open-source Java viewer for chemical structures in 3D. http://www.jmol.org, 2008.

[KJ98] G.J. Kleywegt and T.A. Jones: Databases in protein crystallography. *Acta Cryst,* D54:1119-1131. CCP4 Proceedings, 1998.

The photosynthetic apparatus of Rhodobacter sphaeriodes as an example for a comprehensive systemic reconstruction of a metabolic system

Tihamer Geyer*, Xavier Mol, Sarah Blass, Florian Lauck and Volkhard Helms

Center for Bioinformatics, Saarland University, D-66041 Saarbrücken

*Corresponding author Email: tihamer.geyer@bioinformatik.uni-saarland.de

For simulations of metabolic systems, often only a part of the required rate constants and parameters are known. We investigated, how these can be determined consistently with a systemic model setup which incorporates all available information about the system and a simultaneous parameter fitting against a number of different dynamical experiments. For this process we used the simple and well characterized photosynthetic apparatus of the purple bacterium *Rhodobacter sphaeroides*.

During the setup process, the size of the chromatophore vesicles housing the photosynthetic apparatus put restrictions on, e.g., the relative placement and stoichiometries of the proteins and on their connectivity. Vice versa, kinetic constraints aided in the spatial reconstruction of the very small chromatophores, which carry a total of less than a hundred membrane proteins. To account for the fluctuations in this small system, a molecular stochastic simulation was set up, where each protein with their explicit binding sites and internal states is modelled individually. In this bottom-up approach, the model for each of the proteins is constructed from the relatively well understood microscopic charge transfer or association and dissociation reactions, while the overall connectivity is not fixed a priori, but emerges from the dynamic interplay of the different proteins. We then used an evolutionary algorithm to fit the rate constants and system parameters such that a set of eight experiments, which range from quasi steady state situations to fast multi-flash-triggered scenarios, is reproduced best.

The chosen approach with a stochastic simulation together with an evolutionary search for sets of model parameters is a promising route to a systemic modelling and understanding of metabolic systems. These techniques can also be applied to gene regulation or mixed systems of metabolism and regulation. We also discuss possible extensions of the current model of bacterial photosynthesis.

Stabilizing regions in membrane proteins

Frank Dressel*, Annalisa Marsico, Anne Tuukkanen, Rainer Winnenburg, Michael Schroeder and Dirk Labudde

Biotechnology Center, TU Dresden, Tatzberg 47-49, Dresden, Germany

*Corresponding author Email: frank.dressel@biotec.tu-dresden.de

Around one third of a typical genome consists of membrane proteins. Misfolding of membrane proteins can often be linked to diseases, so that it is of great importance to understand, which residues and interactions are crucial for the stability of these proteins. We developed a coarse-grained model to predict stabilizing regions in membrane proteins. We compare the model to experimental data from Single molecule force spectroscopy (SMFS) and literature to evaluate the effects of mutations on function and stability of five membrane proteins (bacteriorhodopsin, halorhodopsin, rhodopsin, an Na+/H+ antiporter, and aquaporin). The aim of this study is to describe all these data in an unified context, the interaction energies of amino acids in a coarse grained model to gain a better understanding of membrane proteins.

Origin of bacterial outer membrane β-barrels by multiple duplication of a $\beta\beta$ hairpin

Michael Remmert, Andreas Biegert, Dirk Linke, Andrei N. Lupas*, and Johannes Söding*

Max-Planck Institute for Developmental Biology, Tübingen, Germany

Protein Bioinformatics and Computational Biology, Gene Center, LMU Munich, Germany

*Corresponding author Email: soeding@lmb.uni-muenchen.de, andrei.lupas@tuebingen.mpg.de

How did today's complex protein domains originate? We believe that protein domains arose as combinations of peptide modules that originally evolved as cofactors in the RNA world [4]. Here, we investigate the hypothesis that the $\beta\beta$ hairpin that forms the repeating structural unit of the outer membrane β-barrels (OMBBs) represents such an ancestral peptide. Despite the obvious structural similarity between OMBBs and the resulting amphipathic character of their sequences, sequence similarity is hardly detectable [2] and evidence for a common ancestry of all bacterial OMBBs has not been found. Three lines of evidence are provided for our hypothesis that OMBBs originated by multiple duplication of an ancestral $\beta\beta$ hairpin. First we show that most bacterial OMBBs may be linked by a novel remote homology detection method [3, 5] to the exclusion of analogous non-membrane β-barrel folds. Second, using a new method for de-novo repeat detection [1], we detect a clear repeat pattern in the sequences of many OMBBs, the repeat unit each time coinciding with the β hairpins. Third, we combine analysis of structural and sequence similarity to show that the observed sequence similarity between OMBBs cannot be explained by structural constraints on the sequence and hence is a sign of their common origin. All three computational approaches thus support the common origin of canonical bacterial OMBBs by sequential duplication of an ancestral $\beta\beta$ hairpin.

References

1. A. Biegert and J. Söding. De novo identification of highly diverged protein repeats by probabilistic consistency. *Bioinformatics*, 24(6):807–814, Mar 2008.

2. A. F. Neuwald, J. S. Liu, and C. E. Lawrence. Gibbs motif sampling: detection of bacterial outer membrane protein repeats. *Protein Science*, 4(8):1618–1632, 1995.

3. J. Söding. Protein homology detection by HMM-HMM comparison. *Bioinformatics*, 21(7):951–960, 2005.

4. J. Söding and A. N. Lupas. More than the sum of their parts: on the evolution of proteins from peptides. *Bioessays*, 25(9):837–846, Sep 2003.

5. J. Söding, M. Remmert, A. Biegert, and A. N. Lupas. HHsenser: exhaustive transitive profile search using HMM-HMM comparison. *Nucleic Acids Research*, 34(Web Server issue):–374, 2006.

Genome wide identification of transmembrane beta barrel proteins based on residue exposure patterns

Sikander Hayat*, Aaron Goldmann and Volkhard Helms*

Saarland University, Saarbrücken, Germany
University of Pennsylvania, Department of Biology, Philadelphia, USA

*Corresponding author Email:s.hayat@bioinformatik.uni-saarland.de

We first compiled a dataset of 23 non-redundant X-ray structures of TMBs, and generated multiple sequence alignments of homologous protein sequences. Based on the frequency profile obtained from these alignments, we derived the amino acid propensities using ridge regression, such that the positional scores are maximally correlated with the relative solvent accessible surface area of each residue [1]. We then implemented a support vector classifier to predict the exposure status of each TM residue. The relative abundance of each amino acid in exposed and buried state was calculated from the dataset and normalized against the E. coli genome. The normalized frequencies were combined with the predicted exposure status to identify TMBs from the E. coli genome. We show that the results obtained by using predicted exposure status of each residue are at least as good as employing dyad repeat patterns, as previously described in the literature [2]. The method can be employed to identify TMBs from a given genome and to predict the TM beta strands. Furthermore, it also provides the exposure status of each TM residue along with a confidence score of the pre-dictions made.

References:

1.Park, Y., Hayat, S. and Helms, V. 2007. Prediction of the burial status of transmembrane residues of helical membrane proteins. BMC Bioinformatics 8:302.

2.Wimley, W.C. 2002. Toward genomic identification of beta-barrel membrane proteins: composition and architecture of known structures. Protein Sci. 11:301–312.

Prediction of a protein channel structure based on stochastic formal grammars and the continuous ion flow model

Witold Dyrka[1], Jean-Christophe Nebel[2], Malgorzata Kotulska[1*]

[1] Institute of Biomedical Engineering and Instrumentation
Wroclaw University of Technology,
Wybrzeze Wyspianskiego 27, 50-370 Wroclaw, Poland

*Malgorzata Kotulska Email: <kotulska@pwr.wroc.pl>

The best results in 3D structure prediction of protein channels have been obtained by homology based methods, yet this approach is restricted by the small number of solved structures. De-novo techniques are not capable of dealing with medium and large molecules of protein channels. However, the prediction of a few characteristic points can extend its usage to larger proteins by constraining conformational search. We propose to use a Stochastic Context Free Grammar (SCFG) based framework to detect certain characteristic points and substructures. In our previous work, we showed the strength of this scheme is that grammar can be induced automatically, using a genetic algorithm, from a set of unrelated protein sequences which share a common feature [1]. The SCFG framework can be adapted for robust detection of characteristic points by appropriate choice of amino-acid properties and by using problem specific production rules and bracketed inputs. Prediction of the structure of ion channels without known homologous proteins can also be improved by comparing functional characteristics obtained from the model with experimental data. The Poisson-Nernst-Planck (PNP) model, although an averaged theory, is capable of reproducing biologically valid characteristics. We have optimised the PNP method in the terms of computational cost and prepared the pipeline for obtaining the current-voltage and conductance-concentration characteristics for a given channel structure [2]. The criterion of compatibility of the experimental and simulated characteristics can be then used to choose the best structural model of the channel or even to provide some feedback into the 3D modelling process.

References:

[1] Dyrka W, Nebel J-C 2008. A Stochastic Context Free Grammar based Framework for Analysis of Protein Sequences (submitted).
[2] Dyrka W., Augousti AT, Kotulska M 2008. Ion flux through membrane channels - an enhanced algorithm for the Poisson-Nernst-Planck model, Journal of Computational Chemistry, 29/12, 1877-1888.

From Structure to Sequence: Discovery of Novel Motifs in Transmembrane Proteins

Annalisa Marsico*, Andreas Henschel, Boris Vassilev, Anne Tuukkanen, Christof Winter, Ivan Popov and Michael Schroeder

Biotec, TU Dresden

*Corresponding author Email: annalisa.marsico@biotec.tu-dresden.de

A large proportion of genomes encodes for membrane proteins, which are important for many cellular processes. Many diseases can be linked to mutations in membrane proteins. With the tremendous growth of sequence data, there is a need to identify membrane proteins from sequence, to functionally annotate them, and correctly predict their topology. We propose a new computational approach that learns sequential motifs from 3D structure fragments of membrane proteins. The algorithm clusters these fragments according to hydrogen bonding patterns, backbone torsion angles, and sequence conservation and then derives a regular expression representing each cluster.

Prediction of helix-helix contacts and interacting helices in polytopic membrane proteins

Angelika Fuchs#, Andreas Kirschner#, and Dmitrij Frishman*>

Department of Genome Oriented Bioinformatics, Technische Universität München, Wissenschaftszentrum Weihenstephan, 85350 Freising, Germany

#Authors contributed equally

*Corresponding author Email: d.frishman@wzw.tum.de

Despite increasing numbers of available 3D structures, membrane proteins, which constitute up to 30% of a genome, still account for less than 1% of all structures in the Protein Data Bank. Additionally, recent high resolution structures indicate a clearly higher structural diversity of membrane proteins than initially anticipated, motivating the development of reliable structure prediction methods for membrane proteins.

A commonly addressed 2D structure prediction problem in soluble proteins is the prediction of residue-residue contacts. Here, we present a newly developed neural network approach to predict helix-helix contacts specifically in α-helical membrane proteins. Input features for this neural network are both input features commonly used for the contact prediction of soluble proteins like windowed residue profiles and residue distance in the sequence, but also features that apply to membrane proteins only, such as a residue's position within the transmembrane segment or its orientation towards the hydro- or lipophilic environment. Trained on a dataset of 62 membrane proteins with solved structure, the obtained neural network can predict contacts between residues in transmembrane segments with nearly 26% accuracy and recalls 3.5% of all helix-helix contacts. It is therefore the first contact predictor developed specifically for α-helical membrane proteins performing with equal accuracy to state-of-the-art contact predictors available for soluble proteins. The predicted helix-helix contacts were employed in a second step to identify interacting helices. For this prediction problem we gained an accuracy of 78.1%, a sensitivity of 53.1% and a specificity of 86.3%. Compared to our recently published method for identifying interacting alpha-helices based on the analysis of co-evolving residues (1), this corresponds to respective improvements of 6.2%, 11.5% and 3.3%.

References

1. Fuchs A., Martin-Galiano, A.J., Kalman, M., Fleishman, S., Ben-Tal, N. and Frishman D. (2007) Co-evolving residues in membrane proteins, *Bioinformatics*, **23**, 3312-3319.

Organization of the membrane protein fold jungle

Sindy Neumann, Holger Hartmann, Antonio J. Martin-Galiano, and Dmitrij Frishman*

Department of Genome Oriented Bioinformatics, Technische Universität München,
Wissenschaftszentrum Weihenstephan, 85350 Freising, Germany

*Corresponding author Email: d.frishman@wzw.tum.de

Significant progress in structure determination techniques allows to overcome some of the difficulties in the experimental solution of membrane protein structures. However, less than 1% of the proteins of known structure in the Protein Data Bank (PDB) are membrane proteins, although they constitute 20-30% of the proteins in a genome. Thus, bioinformatics methods are needed to narrow the large gap between the number of available sequences and known structures.

Here, we present a revised version of the CAMPS database (1) that presents an automatic classification of membrane proteins into fold classes. Compared to the previous version, CAMPS now covers also sequences from eukaryotic genomes and includes an improved definition of a membrane protein fold. First of all, an initial dataset of 286,476 membrane proteins with at least three predicted transmembrane helices from 535 prokaryotic and eukaryotic genomes was subjected to markov clustering. Out of these initial clusters, we selected those that are homogeneous in terms of sequence identity, the number of predicted transmembrane helices and loop length patterns. Using this procedure we obtained a set of 1384 clusters covering more than 65% of the initial dataset that are believed to represent membrane protein folds. We found that 65 of these clusters already have associated known structures which implies that 1319 additional new structures would be necessary to guarantee complete knowledge of the existing membrane protein fold space.

The CAMPS database thus provides a rough organization of the membrane protein sequence space that is established fully automatic. Its main application lies in the target selection for structural genomics of membrane proteins that approaches to get at least one structural representative for every fold. Increasing the number of representatives results in an increased library of distinct folds which in turn improves the application of structure prediction methods. The database may be further used as a reference for studies on membrane protein evolution, mutations and disease. The latter analyses are of particular interest as membrane proteins are of great pharmaceutical importance.

References

1. Martin-Galiano, A.J. and Frishman, D. (2006) Defining the Fold Space of Membrane Proteins: the CAMPS Database. *Proteins*, **64**: 906-922

The Modular Structure of Cytochrome P450 Monooxygenases

Demet Sirim, Florian Wagner and Jürgen Pleiss*

Institute of Technical Biochemistry, University of Stuttgart
Allmandring 31, 70569 Stuttgart, Germany

*Corresponding author Email: Juergen.Pleiss@itb.uni-stuttgart.de

Cytochrome P450 monooxygenases (CYPs) are ubiquitous heme-containing enzymes which catalyze a wide variety of oxidative reactions. This monooxygenation reaction requires a redox partner to transfer electrons which are essential for binding and activation of the iron-bound molecular oxygen. Using comparative sequence and structure analysis, protein family-specific parameters can be derived to understand the molecular function. To perform this systematic analysis the integration of data on sequence, structure and function is necessary. Therefore we established the Cytochrome P450 Engineering Database (CYPED). The current version is available at www.cyped.uni-stuttgart.de and contains beside more than 6000 sequences about 30 crystal structures.

Despite their differences in sequence and substrate specificity, the structures of CYPs are highly similar: CYPs consists of conserved regions that are essential for structure and function, and of variable regions that mediate the individual biochemical properties. Since the structures are so clearly related, they are compared in detail to define a common core and to assign functions to variable regions. The superposition of structures and generation of a corresponding structure-based sequence alignment is an essential step of such an analysis. A systematic comparison of CYPs that involves sequence and structure should provide a basis to analyze the modular structure of CYPs, to derive biochemical properties: by substrate specificity, selectivity, and interaction with reductase.

In Silico Screening of Drug-Like Compounds Online: eDrugScan

Oliver Frings and Michael C. Hutter *

Center for Bioinformatics, Building C 7.1, Saarland University, Germany

*Michael C. Hutter Email: michael.hutter@bioinformatik.uni-saarland.de

Selecting potentially suitable compounds for experimental testing from the vast chemical space is still a challenge in computer-aided drug design. Corresponding prediction methods comprise individual ADME properties as well as drug-likeness criteria and indices.[1] We have investigated the suitability of decision trees and support vector machines for the classification of chemical compounds into drugs and nondrugs.[2] To account for the requirements upon screening virtual compound libraries, schemes for successive filtering steps with gradual increasing cost were derived. We found that a decision tree approach that uses a minimum of rapidly computable descriptors including Hutter's drug-likeliness index,[1] molar refractivity, molecular weight, and XlogP is most efficient for this purpose.[2] Together with other drug-likeness criteria this filtering scheme has been included in the online tool eDrugScan.[3] To also enable customized step-wise screening including other criteria such as SMARTS provided by the user, the sequence of the filter modules can be arranged interactively. They also allow to specify upper and lower margins for a series of descriptors such as molecular weight, XlogP, and number of rotatable bonds. Currently, eDrugScan accepts uploaded compounds in the .hin file format of HYPERCHEM.[4]

Welcome at eDrugSCAN
Main | Modules | Screening | Results | Compounds | Statistics | Help | FAQ

[1] M.C. Hutter, *J. Chem. Inf. Model.*, **2007**, *47*, 186-104.
[2] N. Schneider, C. Jäckels, C. Andres, M.C. Hutter, *J. Chem. Inf. Model.*, **2008**, *48*, 613-628.
[3] http://service.bioinformatik.uni-saarland.de/edrugscan/
[4] HYPERCHEM, HyperCube Inc, Gainsville, FL, http://www.hyper.com

Bioisosteric Similarity of Molecules Based on Structural Alignment and Observed Chemical Replacements in Drugs

Markus Krier and Michael C. Hutter *

Center for Bioinformatics, Building C 7.1, Saarland University, Germany

*Michael C. Hutter Email: michael.hutter@bioinformatik.uni-saarland.de

Choosing compounds for screening is difficult problem due the vast chemical space. The question is thus where to start. The empirical knowledge of medicinal chemists tells us that similar compounds are likely to have similar properties, which is reflected by so-called bioisosteric replacements.[1] These comprise simple exchanges of terminal atoms as well as more complex structural modifications, such as ring closures. To detect and evaluate all kinds of replacements we have designed an approach that adopts the algorithmic concept used to assess the homology of amino acid sequences to chemical molecules. The mutual exchange frequencies between distinct atom types are expressed in a substitution matrix.[2] Likewise, a pair-wise alignment between the molecules is constructed using dynamic programming.[3] To obtain the actual exchange frequencies, we refined an initial matrix based on observed chemical replacements[4] by collecting the generated alignments of substances from 33 drug classes in an automated procedure. To compute the mutual bioisosteric similarity between molecules a specific function has been derived.[5]

Results

The bioisosteric similarity can be used to express the chemical diversity within a given compound class, e.g. inhibitors of the HIV Reverse Transcriptase are more divers than Angiotensin-II Antagonists. Furthermore the suitability for virtual screening was investigated. For this purpose we compared the recovery of known drugs against a background of other substances from various databases. The majority of drugs possess a higher similarity within the same class than compared to randomly chosen substances from either ZINC[6] or ChemBank.[7] Moreover, nondrugs without any pharmaceutical functions exhibit considerably lower similarities.

[1] G.A. Patani, E.J. LaVoie, *Chem. Rev.* **1996**, *96*, 3147.

[2] S. Henikoff, J.G. Henikoff, *Proc. Natl. Acad. Sci. U.S.A.* **1992**, *89*, 10195.

[3] T.F. Smith, M.S. Waterman, *J. Mol. Biol.* **1981**, *147*, 195.

[4] R.P. Sheridan, *J. Chem. Inf. Comput. Sci.* **2002**, *42*, 103.

[5] M. Krier, M.C. Hutter, *submitted for publication*.

[6] http://zinc.docking.org

[7] http://chembank.broad.havard.edu

Discovery of novel chemokines with automated structure-based protein annotation methods

Aurelie Tomczak* and Maria Teresa Pisabarro

BIOTEChnology Center of the TU Dresden, Dresden, Germany

*Corresponding author Email: arelie.tomczak@biotec.tu-dresden.de

Background and aims

Classic bioinformatics approaches based on sequence similarity are often useful to find homologous proteins and thus infer the function of unknown ones. In case sequence-based annotation methods fail, the application of structure-based methods (i.e. Fold recognition) can provide functional clues or confirm tentative functional assignments. These methods have already proven to be successful for single proteins but, due to the large amount of data obtained by sequencing and other high throughput approaches, automation of structure-based functional annotation of proteins is needed.

Or aim has been to develop a framework for automatic structure-based annotation of proteins and use it to find novel chemokines, which are small secreted signal proteins playing an important role in immune response by guiding immune cell migration. We integrated data and tools from different sources: public databases containing experimental structural and sequence data (i.e. PDB, Swiss-Prot), sequence feature prediction tools (signal peptide, trans-membrane, secondary structure prediction), structure-based computational methods (i.e. Fold recognition) and in-house developed methods describing structural and functional features of the chemokine fold family (3D-descriptors of disulfide bonds and fold coverage). We extracted not annotated protein sequences containing cysteine residues from public databases and used them as input for our framework to discover novel chemokines with remote sequence identity.

Results and outlook

So far we have screened non-annotated human protein sequences and found several candidates for putative chemokines. Detailed 3D models of these protein sequences with potential to be novel chemokines have been built. Their stability and energetic properties have been characterized using computational techniques such as molecular dynamics simulations and promising candidates are currently in the process of being tested experimentally for chemotactic activity.

Design of MHC binding peptides by using Ant Colony Optimization

Paul Wrede*, Natalie Jäger, Jan Hiss, Joanna Wisniewska, Peter Walden, Gisbert Schneider

Institut für Molekularbiologie und Bioinformatik, Charité-Universitätsmedizin Berlin, 14195 Berlin

*Corresponding author Email: paul.wrede@charite.de

Introduction

Optimization of MHC I binding peptides is a substantial support for the development of peptide vaccines. We developed a concept for the design of MHC I stabilizing peptides by combining an Ant Colony Optimization (ACO; Dorigo *et al.*, 1996) algorithm with machine learning methods, artificial neural networks (ANNs) and support vector machines (SVMs), respectively. The machine learning methods serve as fitness functions by classifying octapeptides that are recognized by mouse MHC I allomorph H-2K^b into two categories: MHC I stabilizing and MHC I non-stabilizing.
MHC I proteins are integral cell-membrane proteins which present short peptides with an average length of eight to nine residues. Peptide binding to MHC I molecules stabilizes the MHC/peptide complex at the cell surface, which is a necessary condition for triggering an adaptive immune response (Townsend *et al.*, 1986). The sequences of MHC I binding peptides depend on the genetic haplotype of MHC I. Each haplotype recognizes a prevailing sequence pattern. Studies of more than ten years led to the identification of a canonical sequence pattern for a given haplotype (Rammensee *et al.*, 1995). Besides generating peptides with this common pattern, our study also aimed to circumvent the canonical sequence pattern in H-2K^b binding octapeptides during the optimization process. Although the human cellular immune system is different from that of the mouse we used murine H-2K^b haplotype cells for testing the designed MHC I binding peptides, due to the broad and readily available sequence data (databases AntiJen, IEDB) and readily available *in vitro* assays.

ACO represents an instance of Artificial Intelligence: the swarm intelligence. Swarm intelligence strongly relies on self-organization which allows the highly coordinated behavior of a social insect society. Self-organizing particle systems, like ant colonies consist of numerous autonomous, purely reflexive agents whose collective movements through space are determined primarily by local influences. The size of the ant colony and the representation of the ant's search space is unique referring to the given optimization problem. We demonstrate how the performance of the implemented ACO algorithm depends on the colony size and the size of the search space.
In this study, we focus on the ACO parameters important for the design of MHC I binding peptides, and provide algorithmic details along with parameter optimization. Furthermore, we present several novel MHC I binding - as well as non-binding - peptides not obeying the known sequence motif, which were identified by our ACO algorithm combined with machine learning methods.

Ant Colony Optimization

Ant Colony Optimization (ACO), as first proposed by Dorigo *et al.* in 1996 is a metaheuristic that utilizes an analogue of ant trail pheromones to solve combinatorial optimization problems. Ants find the shortest path between the nest and food source by the use of probabilistic rules based on local information, i.e. chemical substances called pheromones. Artificial ants can do similarly by using also the local information, i.e. artificial pheromones. The intrinsic optimization procedure facilitates a guided walk through the sequence space. In this study, guidance was realized by a specifically trained jury artificial network system.

Output values of five with different descriptor sets trained artificial neural networks served as input for a jury network (Schneider and Wrede, 1998). The training data set consisted of octamer sequences, which were evaluated for their binding property of MHC I allomorph H-2K^b. Several free parameters of the ACO can be varied and optimized to gain a reliable instrument for peptide design.

Which parameters can be varied? First, the number of ants (agents) exploring the sequence space can be varied, reasonably is the use of 1 to 10 to keep computation time low. Second, the representation of ant's search space can be altered. In using the 'trail laying-trail following' metaphor for optimization purposes, the most crucial step is the problem-specific representation of these pheromone trails, which is the artificial ants' search space. In ACO, the pheromone representation is associated with the solution components used by the ants to construct new solutions – in our case, the transition probability between every two amino acids of the octapeptide translate into a pheromone concentration ranging from zero to one in the ants' search space.

Peptide design by ACO

The ACO algorithm was implemented for the systematic navigation through the 20^8 sequence space, which yielded the design of octapeptides. The ants search space, here termed 'pheromone matrix' was constructed to accomodate the given optimization problem of generating new MHC I stabilizing octapeptides. Since there are 20 possible amino acids for each position in the peptide and eight positions to occupy, the size of the pheromone matrix is 20x8. This matrix stores the pheromone concentrations, which represent transition probabilities between two neighboring amino acids in the octapeptide. Thus a peptide was regarded as a path of an ant in the pheromone matrix.

The implemented ACO algorithm consists of three steps:

1. sequence design: realized by an ant moving through the pheromone matrix
2. path evaluation: the path of an ant through the pheromone matrix is evaluated by a fitness function (machine learning method)
3. pheromone update: the pheromone concentrations at the residue positions along the path representing the evaluated peptide are updated according to the fitness function

The loop of these three consecutive steps represents a single iteration of the implemented ACO algorithm. The ACO was iterated until a convergence criterion was reached (each amino acid in the generated peptide has a steady certain pheromone concentration for at least 10,000 iterations).

References

Hiss JA, Bredenbeck A, Losch FO, Wrede P, Walden P, Schneider G. (2007) Design of MHC I stabilizing peptides by agent-based exploration of sequence space. Protein Eng Des Sel. 20(3):99-108.

Rammensee HG, Friede T, Stevanoviíc S. (1995) MHC ligands and peptide motifs: first listing. Immunogenetics. 41(4):178-228.

Dorigo, M., Maniezzo, V., Colorni, A. (1996). Ant system: optimization by a colony of cooperating agents. IEEE Transactions on Systems, Man, and Cybernetics-Part B. 26: 29-41.

Townsend AR, Rothbard J, Gotch FM, Bahadur G, Wraith D, McMichael AJ. (1986) The epitopes of influenza nucleoprotein recognized by cytotoxic T lymphocytes can be defined with short synthetic peptides. Cell.; 44(6):959-68.

Schneider G, Wrede P (1998): Artificial neural networks for computer-based molecular design. Prog. *Biophys. Mol. Biol.* **70**: 175-222.

Structural modeling of a-galactosidase and a-galactosaminidase from Aspergillus Niger

Natallia Kulik*, Lenka Weignerová, Karel Bezouška, Vladimir Křen and Rüdiger Ettrich

LHPC, Institute of Physical Biology, University of South Bohemia, Nové Hrady, Czech Republik *Corresponding author Email: kulik@greentech.cz

Alpha-galactosidases are enzymes (EC 3.2.1.22) which occur widely in microorganisms, plants, and animals, and catalyze the hydrolysis of 1,6-linked α-galactose residues from oligosaccharides and polymeric galactomannans.

A large screening study of extracellular α-N-acetylgalactosaminidase activity of a library of filamentous fungi (42 strains) led to the identification of the best constitutive producer - Aspergillus niger CCIM K2. The enzyme responsible for the activity was purified, biochemically characterized and N-terminally sequenced and has been identified as galactosidase encoded by gene variant A.

Sequence comparison of the sequences of galactosideses variant A (aglA) and variant B (aglB) shows more than 50% identity to α-galactosidases, and only 33% identity to α-galactosaminidase for the aglB. However, the enzyme encoded by aglA has only a 28% identity to the available crystal structure of α-galactosidase from Trichoderma reesei, but 34% identity to α-galactosaminidase from chicken.

Homology models of both galactosidases from A.niger were built with Modeller9.1 and refined by energy minimization and molecular dynamics in YASARA using the Amber force field. Analysis of models and comparing to the known enzymes revealed significant difference in the size of the active centres in aglA and aglB, which can explain the specificity to hydrolysed carbohydrates. Substrate docking clearly demonstrates the preference of the identified enzyme for α-D-N-acetylgalactosamine over galactose, thus giving evidence of the fact that the α-D-N galactosidase type A gene from Aspergillus niger encodes a fully functional α-N -acetylgalactosaminidase.

Molwind – Mapping Molecule Spaces to Geospatial Worlds

Oliver Karch*, Christian Herhaus, Sebastian Bremm and Friedrich Rippmann

Merck Serono, Bio- and Chemoinformatics, Frankfurter Str. 250, 64273 Darmstadt, Germany

*Corresponding author Email: Oliver.Karch@merck.de

Visualizing molecular contexts represented by genes, proteins, bioactive compounds etc. and their relationships remains a challenging task. Metabolomics and High-throughput-screening (HTS) are producing potentially large sets of molecular entities which often need to be explored in an interactive way to allow high-dimensional properties of a hierarchical attribute space to be inspected simultaneously. Here we present a novel approach to visualize a molecule space by mapping it to a geospatial world. This enables the interactivity offered by modern geospatial browsers such as NASA's World Wind (NWW) [1] specifically in the area of view-dependent level-of-detail rendering to be leveraged.

Introduction

In drug discovery, identification of promising lead compounds from HTS series analysis usually involves multi-parametric exploration of organized compound spaces. Visual approaches (e.g. [2]) may help to guide the investigation process. In addition, mapping of structural space to geospatial layers can provide unique ways of intuitive navigation (e.g. elevation, level-of-detail etc.) to be applied to data exploration.

Material and Methods

NWW [1] is a browser to navigate interactively on planetary terrain. The NWW server provides image tiles requested by the browser-client depending on the current level-of-detail (layer) and visibility of view [3]. We have implemented a World Wind server which dynamically generates molecule layers and tiles from a set of chemical structures that have been hierarchically partioned e.g. by substructures.

Results

Our server enables scientists to interactively browse the chemical compound space efficiently changing between different levels of structural detail while maintaining relationships between similar (neighboring) compound classes. Conceptually, it allows arbitrarily layered data-sets to be served to the NWW client and can therefore be easily extended to visualize e.g. protein families or pathway modules.

References

[1] NASA World Wind, http://worldwind.arc.nasa.gov/

[2] Schuffenhauer A., Ertl P., Roggo S., Wetzel S., Koch M.A., Waldmann H.: The Scaffold Tree - Visualization of the Scaffold Universe by Hierarchical Scaffold Classification. J. Chem. Inf. Model., 47 (1), 2007: 47 -58.

[3] Bell, D.G., Kuehnel, F., Maxwell, C., Kim, R., Kasraie, K., Gaskins, T., Hogan, P. Coughlan, J: NASA World Wind: Opensource GIS for Mission Operations Tile. Aerospace Conference, 2007 IEEE: 1-9.

A comparative chemogenomics strategy to predict potential drug targets in the metazoan pathogen, Schistosoma mansoni

A. Rohwer*, C. R. Caffrey, F. Oellien, R. J. Marhöfer, S. Braschi, G. Oliveira,

J. H. McKerrow, P. M. Selzer

Intervet Innovation GmbH, Schwabenheim, Germany

*Corresponding author Email: andreas.rohwer@sp.intervet.com

Schistosomiasis is a prevalent and chronic helmintic disease in tropical regions. Treatment and control relies on chemotherapy with just one drug, praziquantel and this reliance is of concern should clinically relevant drug resistance emerge and spread. Therefore, to identify potential target proteins for new avenues of drug discovery we have taken a comparative chemogenomics approach utilizing the putative proteome of *Schistosoma mansoni* as compared to the proteomes of two model organisms, the nematode, *Caenorhabditis elegans* and the fruitfly, *Drosophila melanogaster*. Using the genome comparison software Genlight, two separate *in silico* workflows were implemented to derive a set of parasite proteins for which gene disruption of the orthologs in both the model organisms yielded deleterious phenotypes (e.g., lethal, impairment of motility), *i.e.,* are essential genes/proteins.

Finally 35 *S. mansoni* proteins were identified for which druggable protein homologs exist in the literature and 18 of these were homologous to proteins with 3D structures including co-crystallized ligands with which structure-based drug design approaches can be prosecuted.

Solvent-dependancy of the interfacial activation of lipases: A systematical analysis using multiple molecular dynamics

Sascha Rehm and Jürgen Pleiss*

Institute of Technical Biochmistry, University of Stuttgart
Allmandring 31, 70569 Stuttgart, Germany

*Corresponding author Email: Juergen.Pleiss@itb.uni-stuttgart.de

Lipases belong to the family of serine hydrolases and play an important role in different industrial applications. They are able to catalyze the hydrolysis in water and esterifcation in organic solvents. Lipases share structural elements like the α/β-hydrolase fold and an active center with the catalytic triad. Several lipases cover their active center with a fexible lid to make the catalytic triad unaccessible for the surrounding substrates. These lipases show higher activity at water-oil interfaces, called interfacial activity. For these lipases, two different conformations can be determined, the so-called active (open) and inactive (closed) states, which are believed to be the end of the interfacial activation pathway.

In this study multiple molecular dynamics simulations of both the open and closed structure of *Candida rugosa* lipase, *Thermomyces lanuginosa* lipase and *Rhizomucor miehei* lipase were performed in both water and toluene, with a much longer duration than the studies before (20-25 ns). The results were then systematically analyzed. We were able to show the solvent-dependency of the interfacial activation and observed different kinds of conformational changes.

The molecular dynamics simulations were performed on three different clusters, one cluster with 256 x AMD Opteron 2GHz, connected with Myrinet interface, one cluster with 400 x Intel Xeon 3,2GHz, connected with Infiniband, both provided by the HLRS[1], and our own cluster, consisting of 256 x AMD Athlon MP 1800+, connected with Myrinet interface. The simulations were performed with the program AMBER using 16 CPUs, with a total of 7000 CPU-hours per simulation.

[1] Höchstleistungsrechenzentrum Stuttgart, www.hlrs.de

Clustering protein binding sites of the pdb: A case study using Flavoproteins

Kerstin Koch*, Olivia Doppelt, François Delfaud and Joël Janin

Université Paris Sud
IBBMC, BAT 430
F-91405 Orsay-Cedex

MEDIT SA, 2 rue de Belvedere
F-91120 Paliaseau

*Corresponding author Email: kerstin.koch@u-psud.fr

We cluster protein structures of the protein data bank (*PDB*) according to their similarity of ligand binding sites. Structural similarity of binding sites is important to understand the function of proteins and can be used for structure based drug design. No similarities of the overall structure concerning the whole fold of parts of proteins are taken into account, only local similarities on the surface of the protein independent of the amino acid sequence or 3D domain fold. The MED-SuMo software is used as framework for the detection of similar sites [1]. The proteins are represented by their 3D surface structural-chemical features (SCF), like e.g. hydrogen acceptor/donor. Graphs are build from triangles of SCFs and two graph representations are compared by calculating the MED-SuMo score for matching SCFs [1]. Three steps are necessary. First, the comparison of all the binding sites of a dataset used a pairwise comparison system, second a similarity matrix is computed. Finally, the classification is performed using MCL (Markov Clustering) and DBScan algorithm. The first experiments were done on a testset of flavoproteins ($\sim$ 750 protein structure containing $\sim$ 900 relevant binding sites). For the interpretation of the clusters, the silhouette distance was calculated and different annotations like EC classification, GO-terms and Prosite identifiers were used to find a measure for the homogenity of the clusters. We search for clusters containing proteins from different families/superfamilies or from different architectures by using SCOP and CATH classification. Most of the clusters are homogenous concerning the different annotations, however clusters with proteins from different families/superfamilies and architectures were found. These cases might be very interesting and will be further analysed. The framework will be used in future for the clustering of all larger ligand binding sites within the *PDB*.

References

1. M.Jambon, A. Imbert, G. Deléage and C. Geourjon, (2003), A New Bioinformatic Approach to detect common 3D sites in proteins, *PROTEINS: Structure, Function, and Genetics*, **52**, 137-145

SCOPPI – A Structural Classification of Protein-Protein Interfaces

Christof Winter, Andreas Henschel, Wan Kyu Kim, Gihan Dawelbait, and Michael Schroeder*

Biotechnology Center, Tatzberg 47–51, Technische Universität Dresden, 01307 Dresden, Germany

*Corresponding author Email: ms@biotec.tu-dresden.de

SCOPPI, the Structural Classification of Protein–Protein Interfaces, is a comprehensive database that classifies and annotates domain–domain interactions found in all known protein structures. SCOPPI applies SCOP domain definitions and a distance criterion to determine inter-domain interfaces. Interfaces are clustered using geometrical properties and thus classified.

Here, we present several applications for SCOPPI. First, we can predict and model an interaction between two proteins consistently found deregulated in patients with pancreas cancer: the transmembrane protease serine 4, TMPRSS4, and the tissue factor pathway inhibitor 2, TFPI2. Using a SCOPPI structural template, we propose a putative inhibition of TMPRSS4, which is up-regulated in pancreas cancer, by TFPI2, which is down-regulated in pancreas cancer. TMPRSS4 is known to be involved in tissue invasion and metastasis. Second, we show diversity of binding orientation of two proteins domains. Long-chain cytokines, which are part of human cytokines, human growth hormone and human prolactin, can interact with the fibronectin type III domains of their receptors in at least ten different binding orientations. Third, we screen gene fusion events of protein complex subunits for conservation of the binding orientation of the fused proteins. We find a conserved orientation in two out of three cases. Last, we show how the Baculovirus protein p35 mimics the binding site of human inhibitor of apopotosis in order to bind to human caspase. Eventually, this can help the virus to prevent apoptosis by caspase inhibition. Although structurally similiar, there is no apparent sequence similarity present at the binding site, displaying a fine example of convergent evolution.

Encoding the sequence information of SCOPPI interfaces in Hidden Markov Models allows for screening uncharacterised genome sequences and predicting protein binding as well as ligand binding sites.

SCOPPI is online at http://www.scoppi.org for browsing and querying, and available for download upon request.

Proline – tryptophan interactions: protein structure stabilization and molecular recognition

Lada Biedermannová*, Jiří Vondrášek and Pavel Hobza

Center for Biomolecules and Complex Molecular Systems, Institute of Organic Chemistry and Biochemistry, Flemingovo nam. 2,160 00, Prague 6, Czech Republic
And
Laboratory of Ligand Engineering, Institute of Biotechnology, Videnska 1083, 142 20 Prague 4, Czech Republic

*Corresponding author Email: lada.biedermannova@img.cas.cz

Among the twenty standard amino acids, proline has unique properties due to the distinctive cyclic structure of its sidechain. Not only does this cyclic arrangement confer a great conformational rigidity to proline, but, as we show here, it enables this residue to participate in interesting stacked-like interactions with aromatic residues. These interactions may play important role in the stabilization of folded protein structures as well as in stabilizing protein-protein complexes.

For a detailed study of these interactions, several proline-tryptophan complexes derived from experimental structures of proteins and protein-protein complexes were selected. These complexes were than investigated by computational methods known to properly describe the London dispersion energy. The results showed that the interaction energies in the proline-trypthophan complex can be very large, especially in the stacked-like arrangement, where the two residues are in parallel orientation. Such a strong stabilization (up to ~7 kcal/mol) is rather surprising with respect to the fact that only one interacting residue has aromatic character. The importance of dispersion energy for the stabilization of this arrangement was confirmed by interaction energy decomposition using SAPT (Symmetry Adapted Perturbation Theory) method. Moreover, geometry optimizations carried out for the stacked-like complexes show that the arrangements derived from protein structure are close to their gas phase minimum geometry, suggesting that the protein environment has only a minor effect on the geometry of the interaction.

We conclude that the combination of proline's conformational rigidity with its ability to participate in strong interactions with aromatic residues are the key features responsible for the prominent role of PRM (Proline Rich Motives) in protein stabilization as well as recognition processes. [1]

[1] L. Biedermannova, K. E. Riley, K. Berka, P. Hobza & J. Vondrasek, Phys. Chem. Chem. Phys., accepted (2008).

Basic Strategies for Molecular Docking with Scoring Functions

Gwyn Skone*, Stephen Cameron, and Irina Voiculescu

Oxford University Computing Laboratory
Wolfson Building
Parks Road
Oxford
OX1 3QD
UK

*Corresponding author Email: gwyn.skone@keble.ox.ac.uk

In recent years, the potential benefits from high-throughput virtual screening to the drug discovery community have been recognized, bringing an increase in the number of tools developed for this purpose. These programs have to process large quantities of data, searching for an optimal solution in a vast combinatorial range. This is particularly the case for protein-ligand docking [Ha02]. Proteins are sophisticated structures, with complicated ligand-interactions for which either molecule might reshape itself [SC07, Te03]. Even the very limited flexibility model of rigid conformation lists requires a 7-dimensional exploration, and the functions for evaluating pose suitability can be quite complex to calculate [Wa06].

This work brings a pure computer science approach to the field, hoping to improve the speed and accuracy of such tools by exploring principles of function evaluation in the context of an existing commercial docking program. We present some early results from this project, demonstrating a substantial reduction in run time and some improvement to ligand placement.

Acknowledgements

This work is funded by InhibOx, Ltd.

References

[Ha02] Halperin, I., Ma, B., Wolfson, H., & Nussinov, R. Principles of Docking: An Overview of Search Algorithms and a Guide to Scoring Functions. *Proteins*, 2002, *47*, 409–443.

[SC07] Skone, G.S., & Cameron, S.A. Protein Structure Computation. *In: Proc. FBIT.* 2007.

[Te03] Teodoro, M.L., & Kavraki, L.E. Conformational Flexibility Models for the Receptor in Structure Based Drug Design. *Curr. Pharm. Des.* 2003, *9*, 1635–1648.

[Wa06] Warren, G.L., Webster Andrews, C., Capelli, A-M., Clarke, B., LaLonde, J., Lambert, M.H., Lindvall, M., Nevins, N., Semus, S.F., Senger, S., Tedesco, G., Wall, I.D., Woolven, J.M., Peishoff, C.E., & Head, M.S. A Critical Assessment of Docking Programs and Scoring Functions. *J. Med. Chem.* 2006, *49*, 5912–5931.

Scanning Full Protein Surfaces For Inducible Pockets: Discriminating Between True And False Positives

Susanne Eyrisch, Volkhard Helms*

Center for Bioinformatics, Saarland University, P.O. Box 151150, D-66041 Saarbruecken, Germany

*Corresponding author Email: volkhard.helms@bioinformatik.uni-saarland.de

When ligands bind to the protein surface instead of deep pockets the in-silico prediction of their binding sites becomes challenging. We developed two programs named PocketScanner and PocketBuilder for scanning a user-defined region of the protein surface for positions of putative pockets [1]. Depending on the size of this region, the number of inducible pockets may become very large and thus identifying the native binding site may be difficult. Therefore, we aim for developing a strategy for disciminating the true positives (i. e. pockets opening at the native binding site) from the false positives. To this end, the entire protein surface of 10 proteins was scanned for potential pocket positions using PocketScanner. At each surface position on a cubic grid a pocket was induced by minimising the protein energetically in the presence of a generic pocket sphere of 2 Å radius. As long as the internal protein energy did not deteriorate by more than 10% the radius was increased by 1 Å. The maximal possible radius of an induced pocket, the burial count before and after inducing a pocket, the change of the internal protein energy, and the number of potential pocket positions in the vicinity were saved. Furthermore, the number of topological constraints of the protein atoms surrounding the potential pocket position was calculated by tCONCOORD [2]. We tested different classification algorithms using cross validation and found that alternating decision trees and naive Bayes classifiers were best suited for maximising the true positive rate while minimising the false positive rate. Our analysis reveals that on average inducing larger pockets is energetically more favourable at the native binding site. Additionally, this finding is in accordance with our observation that this region is generally less topologically constrained than the rest of the protein surface.

References

1. S. Eyrisch and V. Helms (2008). Designing Binding Pockets on Protein Surfaces using the A* Algorithm. German Conference on Bioinformatics.

2. D. Seeliger, J. Haas, B.L. de Groot (2007). Geometry-Based Sampling of Conformational Transitions in Proteins. Structure 15:1482-1492.

Improvement of Automated Structure Prediction

Andrea Hildebrand*, Michael Remmert, Andreas Biegert and Johannes Söding

Gene Center, LMU Munich, 81377 Munich, Germany

*Corresponding author   Email: hildebrand@lmb.uni-muenchen.de

Knowledge of protein structure is crucial for a deeper understanding of biological function. Automatic structure prediction can help to compute a 3D model for proteins with no experimentally determined structure. This work introduces an enhanced version of the HHpred structure prediction server [1] which scored second place in CASP7. We achieve significant improvements through an optimal template selection strategy, which includes RANKING, FILTERING and MULTIPLE TEMPLATE SELECTION.

- RANKING:
  For optimal template selection we include meaningful information provided by the alignments between query and templates (e.g. secondary structure score). We pick those templates with the highest expected model quality score which is calculated by a regression method acting on various alignment scores.
- FILTERING:
  Automatically built alignments are often too diverse for ranking as one can hardly distinguish between close and distant homologs. We construct profiles from PSI-BLAST alignments by a gradient filtering procedure through all levels of diversity. We then select the optimal template and filter strength by the previously described ranking method.
- MULTIPLE TEMPLATE SELECTION:
  Since the use of more than one template can significantly improve model quality, we developed a strategy that selects the optimal set of templates for multiple template modeling.

Results

On a benchmark set of 507 sequences sampled from the Protein Data Bank (PDB) with a similar distribution of sequence identities as in CASP7 we could show 5% improvement in automated structure prediction of our enhanced server compared to the basic version of HHpred.

References

1. J. Söding, A. Biegert, and A. N. Lupas. The HHpred interactive server for protein homology detection and structure prediction. *Nucleic Acids Res*, 33(Web Server issue), July 2005.

HMMerThread Database: a resource for remotely conserved domains

Bianca Habermann[*], Mathias Müller, Robert Henschel, Benjamin Sohn and Charles Bradshaw

Scionics c/o Max-Planck-Institute for molecular Cell Biology and Genetics

*Corresponding author Email: habermann@mpi-cbg.de

We present the HMMerThread database, a resource for conserved domains with remote sequence conservation. The HMMerThread database is based on the HMMerThread tool that combines sensitive sequence based domain searches with fold recognition in order to detect conserved domains that are within the so-called twilight zone of sequence conservation. We have run HMMerThread searches on 7 model organisms and the human genome to detected weakly conserved domains in these species. The identified weak conserved domain hits were validated through the use of orthologues in other organisms. As an example, the human genome was validated against 3 other species (M. musculus, C. lupus familiaris and G. gallus) and revealed a core set of 1,180 domains that were present in all 4 species and conformed to the expected fold for the predicted domain. The results from human along with 7 other genome-wide HMMerThread searches are presented in a relational database, which also includes comprehensive annotation for all proteins along with a live search feature, allowing the user to perform HMMerThread searches against their own sequences which are not present in the database. Furthermore, conservation between species is shown by an extensive map of othologues, which the user can browse through. Finally, we provide interaction data for conserved domains based on crystal structures by extracting data from the iPFAM database and by combining low-throughput interaction data (HPRD) and iPFAM interactions, we are able to predict potential interaction sites based on the observed conserved domains in order to aid in experimental design.

Population interaction dynamics and evolutionary barriers

P. Reuter*, and A. Torda

Zentrum für Bioinformatik,
Universität Hamburg

*Corresponding author Email: reuter@zbh.uni-hamburg.de

We are interested in modelling the development or evolution of interacting populations. The populations interact via differential equations and the evolution of the whole system is represented by changes in the kinetic parameters of the interactions.

Earlier, we investigated simple theoretical models of populations interacting in a predator-prey relationship. The size of such populations frequently exhibits oscillatory behaviour as a regular part of their lifecycles. It is also affected by (usually detrimental) events that occur randomly.

As evolutionary changes occur on a time scale that is not realistically tractable with population dynamics methods, we are interested in approximating the long term behaviour of simulated populations with respect to stability. This is achieved by assigning a measure of resilience against extinction by random events to a given simulation.

For many parameter configurations, the simulation produces unstable dynamics that are likely to lead to extinction. However, there exist contiguous regions of parameter space with stable behaviour. Within these stable regions, considerable variation in the specific type of behaviour exists, as well as areas of unstable dynamics. If movement in parameter space is constrained to stable regions, parameter configurations within short euclidean distance of each other may become separated by large distances.

Evolutionary trajectories in such a system can easily become constrained to specific types of interaction dynamics, requiring a narrow set of conditions to switch to a different type of interaction. In allopatric speciation, this kind of constraint may contribute as an isolating mechanism.

In these simple systems, it now seems possible to give a probabilistic description of why one sees certain kinds of cooperation and predation, and what changes to this interactions are permitted.

Protein Structure Reconstruction from a 1D-Representation

Katrin Wolff, Michele Vendruscolo, and Markus Porto*

Institut für Festkörperphysik, TU Darmstadt, Hochschulstraße 8, 64289 Darmstadt, Germany

*Corresponding author Email: porto@fkp.tu-darmstadt.de

We present reconstruction of small protein structures from their one-dimensional structure profile [1]. This approach differs from well-known Gō-models in two respects: Firstly, the structure information is encoded in a 1D vector similar to the encoding of structure by sequence. Secondly, the interactions between residues are not of a simple pairwise nature where residues that form native contacts attract each other specifically. The definition of the structure profile rather ensures that there is interaction between *all* residues and native contacts are favoured indirectly. Here, we use profiles computed from known target proteins to reconstruct the native structure but the strong correlation between hydrophobicity and structure profile [2] may also open the possibilty of structure prediction along this pathway. As underlying model in our reconstructions we use a simplified protein representation consisting of a C_α-trace inscribed by a tube of finite thickness to account for steric exclusion. This description is homogenous, sequence/structure information only enters through the structure profile. Following this approach, we are able to reconstruct several small proteins to the same accuracy as in Gō-models based on C_α-contacts and the resulting dynamics reveal an energy landscape that is more realistic than the one encountered in Gō-models.

References

1. Porto, M., Bastolla, U., Roman, H. E., Vendruscolo, M., Phys. Rev. Lett. **92**, 218101 (2004). Wolff, K., Vendruscolo, M., Porto, M.. Gene (2008, in print). doi: 10.1016/j.gene.2008.06.004.

2. Bastolla, U., Porto, M., Roman, H.E., Vendruscolo, M. 2005. Proteins **58**: 22 – 30 (2005). Bastolla, U., Ortíz, A.R., Porto, M., Teichert, F. Proteins (2008, in print). doi: 10.1002/prot.22113.

Mutation tagging with gene identifiers applied to membrane protein stability prediction

Rainer Winnenburg*, Conrad Plake, Frank Dressel, Dirk Labudde, and Michael Schroeder

Biotechnology Center, Technische Universität Dresden, Tatzberg 47-49, Dresden, Germany

*Corresponding author Email: rainer.winnenburg@biotec.tu-dresden.de

The automated retrieval and integration of information about protein point mutations in combination with structure, domain and interaction data from literature and databases promises to be a valuable approach to study structure-function relationships in biomedical data sets.

As a prerequisite, we developed a rule- and regular expression-based protein point mutation retrieval pipeline for PubMed abstracts, which shows an F-measure of 87% for the pure mutation retrieval task on a benchmark dataset. In order to link mutations to their proteins, we utilised a named entity recognition algorithm for the identification of gene names co-occurring in the abstract, and established links based on sequence checks. We identified more than 10Mio genes/proteins in nearly 3.5Mio abstracts and 260.000 mutations in 80.000 of these abtracts (2.3%). In 52% of cases the identified gene's sequence and the mutation are consistent. We evaluated the use of mutations in gene identification in detail on a small test set of 22 abstracts. Identifying the correct gene improved from 77% to 91% when considering the mutations.

To demonstrate practical relevance, we set up a mutation screening for five membrane proteins from the family of G protein-coupled receptors to evaluate a solvation energy based model for the prediction of stabilising regions in membrane proteins. We identified 35 mutations in text. 25 out of 35 mutation phenotypes reported in literature were in compliance with the prediction of the energy model, which supports a relation between mutations and stability issues in membrane proteins.

PoPMuSiC v2.0: Prediction of Protein Mutant Stability Changes using statistical potentials and neural networks

Yves Dehouck*, Aline Grosfils, Benjamin Folch, Dimitri Gilis, Jean Marc Kwasigroch, Philippe Bogaerts, and Marianne Rooman

Unité de Bioinformatique Génomique et Structurale, Université Libre de Bruxelles, CP 165/61, Av. F. Roosevelt 50, 1050 Brussels, Belgium.

*Corresponding author Email: ydehouck@ulb.ac.be

The ability to tune certain physicochemical or biological properties of proteins, through amino acid substitutions, would be very rewarding in various industrial applications. Such mutations may, for instance, increase the protein's solubility, or maintain its activity under unusual pH or temperature conditions. Whatever the modified property, the considered mutations should not alter the protein structure and stability too much, as it could lead to the loss of the main protein function. Since the experimental determination of the stability change upon mutation is time consuming, efficient predictive methods are needed.

Some of us previously developed the PoPMuSiC program (Prediction of Protein Mutant Stability Changes) [1], which was shown to perform quite well, both in preliminary tests [2] and in subsequent blind predictions [3]. Our purpose here is to improve the predictive abilities of this program, which still suffered from certain limitations.

For that purpose, we exploit newly developed statistical energy functions [4], based on a formalism that highlights the coupling between 4 different protein descriptors, as well as the volume variation of the mutated amino acid. The stability change is expressed as a linear combination of these energetic functions, whose proportionality coefficients vary with the solvent-accessibility of the mutated residue. A MultiLayer Perceptron with sigmoid nodes is found to be well suited for identifying these parameters, and allows them to keep a relatively clear biophysical signification. We show that the resulting algorithm presents a significantly improved predictive power, with respect to the previous version of PoPMuSiC, and to other programs described in the literature.

[1] J.M. Kwasigroch et al., Bioinformatics,18:1701-2 (2002).
[2] D. Gilis and M. Rooman, Protein Eng, 13:849-56 (2000).
[3] D. Gilis et al., J Mol Biol, 325:581-9 (2003); D. Gilis et al., Protein Sci, 16:2360-7 (2007).
[4] Y. Dehouck et al., Biophys J, 90:4010-7 (2006).

Translational science: 1 + 1 = 3

H. Venselaar*

Center for Molecular and Biomolecular Informatics, Radboud University Nijmegen

*Corresponding author Email: h.venselaar@cmbi.ru.nl

Today's research requires more and more the cooperation of scientists in different fields. This combination of their strengths and expertise is known as translational science.

Our study combines (bio)medical research and bioinformatics to solve (bio)medical problems using 3D structures.

3D structures are solved by NMR, X-ray crystallography or homology modeling. The latter technique uses a homolog of the protein of interest to make its 3D structure. Compute servers and visualization software are required to obtain, visualize and manipulate the 3D structures. Protein structures are used to explain the effect of pathological mutations, explain binding specificities of ligands and for general insight in the protein of interest. The next step is to use 3D structures for intelligent experimental design. Biomedical answers must be provided in a format that is meaningful to each collaborator; with or without bioinformatic background.

3D structures were used in several collaborative projects. One example is the Leucine to Proline mutation in HFE, found in a clinical study for Hereditary Haematochromatosis. Visualization of the mutation showed that introduction of a proline disturbs a helix at the interaction interface resulting in loss of binding ability to other proteins. This example shows that 3D protein structures are helpful in (bio)medical research.

Collaboration of bioinformatics and medical research is a nice example of the strength of translational science.

CORUM, the Entrance to the Protein Complex Universe

Ruepp*, A., Brauner, B., Dunger, I., Fobo, G., Frishman, G., Montrone, C., Schmidt, T., Waegele, B. and Mewes, H.-W.

Helmholtz Zentrum München - German Research Center for Environmental Health

*Corresponding author Email: andreas.ruepp@helmholtz-muenchen.de

Results

A large fraction of proteins perform their cellular function not as isolated molecules but as members of protein complexes. CORUM (Ruepp et al., 2008) is a database of manually annotated protein complexes from mammalian organisms. All protein complexes in CORUM were experimentally characterized by individual experiments extracted from scientific literature. Thus, intrinsic errors occurring in high-throughput experiments are excluded. To a large extent, information on isolation of complexes and the identity of the subunits is supplemented with functional characterization. Systematic annotation of protein complex functions is performed using the FunCat annotation scheme. With approx. 2500 protein complexes annotated, CORUM is the largest publicly available collection of protein complexes. Subunits of all complexes represent 3000 different genes, which covers 15% of the protein coding genes in e.g. the human genome. Data is available for download as flat file or in the PSI-MI 2.5 XML format.

Reference List:

Ruepp,A., Brauner,B., Dunger-Kaltenbach,I., Frishman,G., Montrone,C., Stransky,M., Waegele,B., Schmidt,T., Doudieu,O.N., Stumpflen,V., and Mewes,H.W. (2008). CORUM: the comprehensive resource of mammalian protein complexes. Nucleic Acids Res. 36, D646-D650.

Evolutionary flexibility of protein complexes

Michael F. Seidl and Jörg Schultz*

*Department of Bioinformatics, Biozentrum, University Würzburg, Am Hubland, 97074 Würzburg, Germany

Corresponding author Email: Jörg Schultz- joerg.schultz@biozentrum.uni-wuerzburg.de

Protein complexes are the main organisational unit of the cell. Due to attachments and shared components, their protein content is dynamic over the lifetime of a cell [1]. In an analysis of the SMN complex, this dynamic was also observed on evolutionary timescale [2] - consecutive addition and secondary losses played a major role in its evolution. Here, we analysed whether this flexibility is an exception or the rule. Using an iterative algorithm for ortholog prediction, we computed interologs for all members of complexes from HPRD [3] in 23 different species. We found that most complexes arose by the consecutive addition of components. Strikingly, secondary losses played a major role in shaping protein complexes. Thus, the composition of protein complexes is also evolutionary highly flexible. Viewing protein complexes as molecular machines, they seem to need an engine but whether power window lifts are necessary or not depends on the owner.

References

1. AC Gavin, et al. Proteome survey reveals modularity of the yeast cell machinery. *Nature*, 440(7084):631–6, Mar 2006.

2. M Kroiss, et al. Evolution of an rnp assembly system: A minimal smn complex facilitates formation of usnrnps in drosophila melanogaster. *Proc Natl Acad Sci USA*, Jul 2008.

3. S Peri, et al. Development of human protein reference database as an initial platform for approaching systems biology in humans. *Genome Res*, 13(10):2363–71, Oct 2003.

Local landscape of protein complexes in the plant *Arabidopsis thaliana*

Hollunder, J.[1]*, Van Leene, J.[1], Eeckhout, D.[1], Stals, H.[1], Buffel, Y.[1], Neirynck, S.[1], Persiau, G.[1], Van Isterdael, G.[1], Van De Slijke, E.[1], Pharazyn, A.[2], Hendricx, K.[2], Laukens, K. [2], Witters, E.[2], Van Onckelen, H.[2], Inzé, D.[1], Kuiper, M.[1], Van de Peer, Y.[1] and De Jaeger, G.[1]*

[1]Functional Proteomics Group, Department of Plant Systems Biology, Flanders Institute for Biotechnology (VIB), Ghent University, Technologiepark 927, B-9052 Gent, Belgium.
[2]Centre for Proteome Analysis and Mass Spectrometry, University of Antwerp, Groenenborgerlaan 171, B-2020 Antwerpen, Belgium

*Corresponding author Email: jehol@psb.ugent.be, gejae@psb.ugent.be

Proteins associate to build protein complexes which act as highly specialized molecular machines, capable to carry out complex cellular processes. There is growing evidence that such protein complexes are composed of hierarchically organized functional modules. In recent years, new technologies have been developed for interactome studies under near-physiological conditions. Especially tandem affinity purification (TAP) (1) combined with mass spectrometry-based protein identification is a powerful approach that recently has led to a genome wide interactome for yeast (2, 3). Similar platforms are now being set up for higher eukaryotic systems. Here, we have set up a high throughput tandem affinity purification/mass spectrometry platform for cell suspension cultures to analyze cell cycle related protein complexes in Arabidopsis thaliana (4). Elucidation of this protein protein interaction network is essential to fully understand the functional differences between the highly redundant cyclin/cyclin dependent kinase modules, which are generally accepted to play a central role in cell cycle control, in all eukaryotes. So far, we identified protein interactions for 120 proteins, known to be involved in the plant cell cycle. Together with data mining methods, such as DASS (5) for extracting 'molecular machines' with different functional modules that contribute to one superimposed cellular task and integration of additional cell-cycle related data, such as phylogenetic data, gene fusion, gene expression, and domain arrangements, the complexity of cell cycle regulation (e.g. G1-S transition regulation and cytokinesis) will be visualized in a protein interactome. The interactome will show an up-to-date picture of components and interactions regarding the cell-cycle in Arabidopsis and give us new insights for fundamental control mechanisms for this pathway, their functionality, resolvability, flexibility, and robustness, as well as new hypotheses about these processes.

(1) Rigaut, G. *et al.* (1999*) Nat. Biotechnol.* 17, 1030-1032
(2) Gavin, A. C. *et al.* (2006) *Nature* 440, 631-636
(3) Krogan, N.J. *et al.* (2006) *Nature* 440, 637-643
(4) Van Leene, J. *et al.* (2007) *Mol. & Cell. Proteomics* 6, 1226-1238
(5) Hollunder, J. *et al.* (2007) *Bioinformatics* 23, 77-83

Clustering by Common Friends Finds Locally Significant Proteins Mediating Modules

Bill Andreopoulos†, Aijun An, Xiaogang Wang, Michalis Faloutsos, Michael Schroeder

†williama@biotec.tu-dresden.de , http://www.proteinclustering.com/

1 Motivation

Much research has been dedicated to large-scale protein interaction networks including the analysis of scale-free topologies, network modules, and the relation of domain-domain to protein-protein interaction networks. Identifying locally significant proteins that mediate the function of modules is still an open problem.

2 Method

We use the MULIC Multiple Layer Incremental Clustering algorithm for interaction networks, which groups proteins by the similarity of their direct neighborhoods (2). We identify locally significant proteins, called mediators, which link different clusters. We apply the algorithm to a yeast network (3).

3 Results

We observed a hierarchy of mediators and clusters. A cluster of proteins is often mediated by a ‘parent’ cluster, and in turn mediates a ‘child’ cluster. We compare the clusters and mediators to known yeast complexes and find agreement with precision of 71% and recall of 61%. We analysed the functions, processes and locations of mediators and clusters. We found that 55% of mediators to a cluster are enriched with a set of diverse processes and locations, often related to translocation of biomolecules. Additionally, 82% of clusters are enriched with one or more functions. The impotant role of mediators is further corroborated by a comparatively higher degree of conservation across genomes.

4 Protein translocation to nucleus

We illustrate the above findings with an example of membrane protein translocation from the cytoplasm to the inner nuclear membrane (1). The most obvious processes that are described by the results involve translocation of 1biomolecules and transmembrane proteins between the cytoplasm and the inner nuclear membrane, through nuclear pores. Translocation of membrane proteins has long been an enigma but it has recently been shown to require energy and nuclear pore complexes (NPCs) together with karyopherins are involved. This translocation is believed to be karyopherin-mediated, through nuclear localization signal binding sites (4). The mediator hierarchy shows proteins’ reliance on receptor-mediated transport through NPCs, as described by Blobel et al. (4). Several mediator dependencies are shown as involved in protein translocation between the cytoplasm and nucleus via NPCs. The top-level mediators include some highly-connected ATPases in the yeast protein interaction network that are involved in releasing energy that drives chemical reactions. The proteins at the lowest levels have more granular functions.

5 Conclusion and Future Work

Proteins with similar interaction partners often comprise a functional module. Cluster involvement in processes is mediated by locally significant mediator proteins that may be of low degree. A cluster may be both a mediator and mediated itself by other cluster(s), resulting in a hierarchy of clusters and mediators. Our cluster-mediator yeast results match the modular complexes of Gavin et al. (3).

References

[1] Bill Andreopoulos, Aijun An, Xiaogang Wang, Michalis Faloutsos and Michael Schroeder. Clustering by common friends finds locally significant proteins mediating modules, Bioinformatics, Oxford University Press, 23(9): 1124-1131, 2007.

[2] Bill Andreopoulos, Aijun An and Xiaogang Wang. Hierarchical Density-based Clustering of Categorical Data and a Simplification. In Proceedings of the 11th Pacific-Asia Conference on Knowledge Discovery and Data Mining (PAKDD 2007), Springer LNCS 4426/2007, pages 11-22, Nanjing, China, May 22-25, 2007.

[3] Gavin, A.C., et al. (2006) Proteome Survey Reveals Modularity of the Yeast Cell Machinery. Nature, 440:30.

[4] King, M.C., Lusk, P.C., Blobel, G. (2006) Karyopherin-mediated import of integral inner nuclear membrane proteins. Nature, 442(7106):1003-7.

Theoretical Models for Graph-Based Data Clustering

Christian Komusiewicz*and Rolf Niedermeier and Johannes Uhlmann*

Institut für Informatik, Friedrich-Schiller-Universität Jena, Ernst-Abbe-Platz 2
D-07743 Jena, Germany.

*Corresponding author Email: {ckomus,uhlmann}@minet.uni-jena.de

We present several theoretical models for graph-based data clustering. These are variations or extensions of the CLUSTER EDITING model, which has been recently applied in clustering protein similarity data [1,4]. In CLUSTER EDITING, relations between data items are modeled in a similarity graph. More specifically, graph vertices one-to-one represent data items, and if two data items have high similarity, they are connected by an edge. The aim is to find a *cluster graph* (here, a disjoint union of cliques) that is closest to the input graph, where "closeness" is measured by the number of edge modifications needed to transform the similarity graph into a cluster graph. The presented models include currently investigated "relaxed" notions of cluster graph (for instance, incomplete or overlapping clusters). We also present extensions to bipartite graphs [2] (as used for representing gene expression data), and a variant allowing vertex deletions instead of edge modifications [3]. We aim to further develop our algorithmic techniques in order to better reflect constraints from practice. This poster intends to invite applied people to cooperate on graph-based data clustering.

Acknowledgements Research supported by Carl-Zeiss-Stiftung and DFG (NI 369/7).

References

1. S. Böcker, S. Briesemeister, Q. B. A. Bui, and A. Truß. A fixed-parameter approach for weighted cluster editing. In *Proc. 6th APBC*, volume 5 of *Series on Advances in Bioinformatics and Computational Biology*, pages 211–220. Imperial College Press, 2008.

2. J. Guo, F. Hüffner, C. Komusiewicz, and Y. Zhang. Improved algorithms for bicluster editing. In *Proc. 5th TAMC*, volume 4978 of *LNCS*, pages 451–462. Springer, 2008.

3. F. Hüffner, C. Komusiewicz, H. Moser, and R. Niedermeier. Fixed-parameter algorithms for cluster vertex deletion. In *Proc. 8th LATIN*, volume 4957 of *LNCS*, pages 711–722. Springer, 2008.

4. T. Wittkop, J. Baumbach, F. P. Lobo, and S. Rahmann. Large scale clustering of protein sequences with FORCE – a layout based heuristic for weighted cluster editing. *BMC Bioinformatics*, 8(1):396, 2007.

Unraveling Protein Networks with Power Graph Analysis

Loic Royer*, Matthias Reimann, Bill Andreopoulos and Michael Schroeder

BIOTEChnology Center of the TU Dresden, Dresden, Germany

*Corresponding author Email: royerloic@gmail.com

Networks play a crucial role in biology and are often used as a way to represent experimental results. Yet, their analysis and representation is still an open problem. Recent experimental and computational progress yields networks of increased size and complexity. There are, for example, small- and large-scale interaction networks, regulatory networks, genetic networks, protein-ligand interaction networks, and homology networks analyzed and published regularly. A common way to access the information in a network is though direct visualization, but this fails as it often just results in "fur balls" from which little insight can be gathered. On the other hand, clustering techniques manage to avoid the problems caused by the large number of nodes and even larger number of edges by coarse-graining the networks and thus abstracting details. But these also fail, since, in fact, much of the biology lies in the details. This work presents a novel methodology for analyzing and representing networks. Power Graphs are a lossless representation of networks, which reduces network complexity by explicitly representing re-occurring network motifs. Moreover, power graphs can be clearly visualized: they compress up to 90% of the edges in biological networks and are applicable to all types of networks such as protein interaction, regulatory networks, or homology networks.

Protein-Protein Interaction Prediction

Florian Fink*, Stephan Ederer and Wolfram Gronwald*

Institute of Functional Genomics, University of Regensburg, Germany

*Corresponding author Email: {florian.fink, wolfram.gronwald}@klinik.uni-regensburg.de

General Idea

Based on a protein-protein docking approach we develop procedures to verify or falsify protein-protein interactions that were proposed by other methods such as yeast2hybrid assays. Our method utilizes intermolecular energies and amino acid based pair-potentials. For the latter we calculate score distributions for protein-protein complexes that exist in nature (native complexes) and those that do not exist (false complexes). A difference in the distributions of the two groups would then open the possibility to discriminate between complexes that exist and those which do not.

Methods

Native complexes are taken from the Nussinov Database[1], false complexes are produced by using the docking algorithm HADDOCK[2] and forcing two non-interacting proteins to build a complex. For both groups intermolecular electrostatic energies, van-der-Waals energies and amino acid based pair-potentials are calculated. Using the so obtained distributions hypothetical complexes can be assigned a probability that they are native.

Results

(a) The obtained distributions for all three scoring functions are actually different in maximum and in shape.
(b) For several complexes we could already show that the intermolecular energies of the native complexes are considerably lower as for the false complexes.
(c) We found a correlation between the RMSD to the native complex and our predicted probabilities for the Barnase-Barstar complex.

References:

[1] Nussinov, R., Non-redundant dataset (http://bioinfo3d.cs.tau.ac.il/Interfaces/Non-Redundant).

[2] Dominguez et al., HADDOCK: A Protein-Protein Docking Approach Based on Biochemical or Biophysical Information JACS, **125**, 1731-1737 (2003).

Linking the spatial organization of protein-protein interactions and metabolic pathways

Pawel Durek* and Dirk Walther

Max Planck Institute of Molecular Plant Physiology, Potsdam, Germany

*Corresponding author Email: durek@mpimp-golm.mpg.de

The study of biological interaction networks is a central theme of systems biology. Here, we investigate the relationship between two distinct types of interaction networks: the metabolic pathway map and the protein-protein interaction (PPI) network. It has long been established that successive enzymatic steps are often catalyzed by physically interacting proteins forming multi-enzymes complexes. Inspecting high-throughput PPI-data, it was shown recently that indeed enzymes involved in successive steps are generally more likely to interact than other protein pairs.

In our study we expanded this line of research to include comparisons of the underlying respective network topologies as well as to investigate whether protein-protein interactions exhibit any apparent evolutionary optimization linking metabolic flux and spatial organization. Using yeast data, we detected long-range correlations between shortest paths in both network types suggesting a relatively close resemblance of both network architectures. Furthermore, enzymes carrying high flux loads are more likely to physically interact than enzymes with lower metabolic throughput. In particular, enzymes associated with catabolic pathways as well as enzymes involved in the biosynthesis of complex molecules show high degrees of physical clustering. Thus, our results may contribute towards revealing unifying principles shaping the evolution of both the functional (metabolic) as well as physical interaction network.

Construction of an integrated knowledgebase for protein-protein interactions

Helena Dierenfeld, Gabriele Petznick, Holger Marquardt, Paul Hammer, Chong Wang, Antje Krause and Peter Beyerlein*

TFH Wildau, University of Applied Science
Bahnhofstr. 1
D-15745 Wildau

*Corresponding author Email: peter.beyerlein@tfh-wildau.de

The understanding of protein interactions is essential for the process of investigating the functions of cells and higher organisms. Beginning from simple protein-protein interactions comprehensive interaction networks can be established and analyzed by means of commercial and non-commercial protein interaction databases. However, these databases show incomplete consistency (even contradictory), which gives uncertain results derived via these databases.

To overcome this problem we integrated some of the most common databases, which gives us the opportunity to evaluate the datasets. Four different protein interaction databases were integrated into one MySQL-Database: IPA (Ingenuity Pathways Analysis - Ingenuity® Systems, www.ingenuity.com) as the only commercial database, MINT [1], Reactome [2] and IntAct [3]. Additionally the databases UniProt [4], GO [5] and PubMed where partly integrated to provide additional information about the analyzed proteins.

As a first approach to analyze the interaction data of the integrated database the mutual information method [6] was established. This method evaluates for all proteins the correlation of the possibility of being a hub while at the same time fulfilling a distinct function.

References

1. Andrew Chatr-aryamontri et al. MINT: the Molecular INTeraction database. *Nucleic Acids Res.*, 35:D572–D574, 2007.

2. G. Joshi-Tope et al. Reactome: a knowledgebase of biological pathways. *Nucleic Acids Res.*, 33:D428–D432, 2005.

3. Henning Hermjakob et al. IntAct: an open source molecular interaction database. *Nucleic Acids Res.*, 32:D452–D455, 2004.

4. The UniProt Consortium. The Universal Protein Resource (UniProt). *Nucleic Acids Res.*, 36:D190–D195, 2008.

5. Gene Ontology Consortium. The Gene Ontology (GO) project in 2006. *Nucleic Acids Res.*, 34:D322–D326, 2006.

6. C. Schmidt et al. Probabilistic analysis of the function of protein interaction hubs. *German Conference on Bioinformatics*, 2007.

Pathway Prediction with eQTL and Gene Interaction Networks

Jacob J. Michaelson, Andreas Beyer*

Biotechnology Center
Tatzberg 47-49
01307 Dresden

*Corresponding author Email: andreas.beyer@biotec.tu-dresden.de

Expression quantitative trait loci (eQTL) have become increasingly popular in recent years as a technique to discover genetic regulators of gene expression. Several problems plague the effective interpretation of eQTL results, including noisy data prone to many false positives, a lack of molecular context for the significant loci, and an inability to identify causal genes within causal loci, due partially to linkage disequilibrium. To address these challenges, we developed a simple network analysis method and applied it to eQTL measurements from mouse BXD strains. The algorithm uses a gene interaction network as a topology, maps eQTL association scores to the genes in the network, and searches for areas of localized score enrichment. These areas of score enrichment are compared to an empirical null distribution to assess significance. The nature of the gene interaction network imparts a molecular context to sets of eQTL, while aiding in filtering non-causative genes from significant loci. To benchmark this network technique, we evaluated eQTL measurements from several tissues from mouse BXD strains. We examined genes whose expression is regulated by well-known pathways, and assessed the ability of our method to recover upstream pathway members, as compared to traditional cut-off approaches to eQTL analysis. Our results indicate a marked improvement over traditional eQTL analysis, and are enriched for known regulatory pathway members.

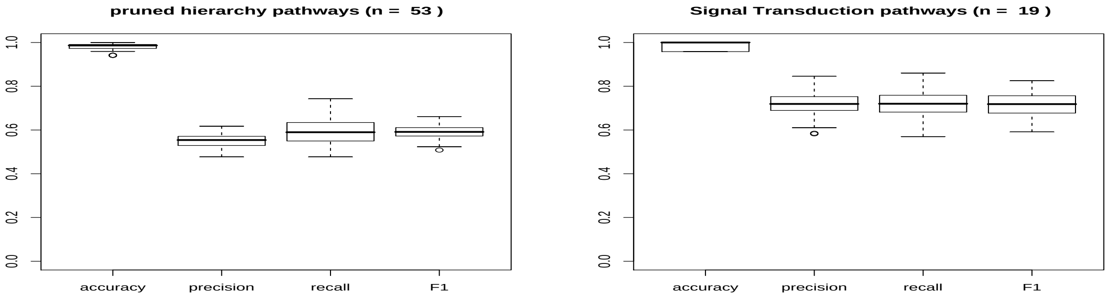


Figure 1: Prediction performance of our method (10×10-fold cross-validation). The accuracy measure uses the same loss-function, which was used to train the classifier, and which takes into account the KEGG hierarchy. **Left:** Pathway prediction within KEGG hierarchy (after pruning the hierarchy for metabolic pathways at the top and for "cellular processes" and "Genetic Information Processing" pathways at the 2nd hierarchy level). **Right:** Pathway component prediction for signaling pathways.

Predicting Pathway Membership via Domain Signatures

Holger Fröhlich*and Tim Beißbarth

German Cancer Research Center (DKFZ), Im Neuenheimer Feld 580, 69221 Heidelberg

*Corresponding author Email: h.froehlich@dkfz-heidelberg.de

Motivation: Functional characterization of genes is of great importance, e.g. in microarray studies. Valueable information for this purpose can be obtained from pathway databases, like KEGG. However, only a small fraction of genes is annotated with pathway information up to now. In contrast, information on contained protein domains can be obtained for a significantly higher number of genes, e.g. from the InterPro database.

Results: We present a classification model, which for a specific gene of interest can predict the mapping to a KEGG pathway, based on its domain signature. The classifier makes explicit use of the hierarchical organization of pathways in the KEGG database. Furthermore, we take into account that a specific gene can be mapped to different pathways at the same time. The classification method produces a scoring of all possible mapping positions of the gene in the KEGG hierarchy. Evaluations of our model, which is a combination of a SVM and ranking perceptron approach, show a high prediction performance. Moreover, for signaling pathways we reveal that it is even possible to forecast accurately the membership to individual pathway components. The complete method is available in the R package *gene2pathway*.

Genetic pathway analysis may point to further applications in of known antineoplastic in additional cancer types.

M. Krupp[1]*, K. Schlamp[1], T. Bauer[2], P. R. Galle[1], A. Teufel[1]

[1]Medical Department, Johannes Gutenberg University, Mainz, Germany
[2]German Cancer Research Center (DKFZ), Heidelberg, Germany

*Corresponding author Email: kruppm@uni-mainz.de

Background

Currently, approximately 1.5 million new cancer cases are expected in the United States in 2008. Beside surgery, chemotherapy including immunotherapy utilizing monoclonal antibodies is among the most commonly applied therapeutic strategies to treat cancer. Although significant progress has been made, the complete potential of those therapies is still to be explored.

Abstract

We analyzed a dataset of 1402 microarrays collected from the Standford MicroArray Database (SMD) [1]. Those data revealed 5922 tumor associated genes, corresponding to 27 distinct cancer types. In addition, 119 antineoplastics and 210 pathway maps were downloaded from the Kyoto Encyclopedia of Genes and Genomes (KEGG) [2]. Finally, antineoplastic application was extracted from several chemical databases. After assigning pathway information to tumor associated genes and antineoplastics, 125 significant pathway maps for cancer progression and 66 pathway maps containing antineoplastic targets were identified. Combining and analyzing these pathway-tumor-antineoplastic relations with respect to differentially expressed gene signatures in these tumor types, we demonstrated a significantly extended potential of individual chemotherapeutics due to a genetic relationship of specific genetic pathways with specific diseases suggesting a potential association of the drug effect on these additional diseases.

Conclusion

Our analysis showed that an analysis of genetic pathways associated with cancer genomics may extend the potential of existing antineoplastics to an efficient application in additional diseases.

References

[1] http://genome-www5.stanford.edu/
[2] http://www.genome.jp/kegg/

Integrating univariate and multivariate statistics with pathway database information for the analysis of microarrays

Frederik Roels*, Ingmar Bruns, Akos Czibere, Benedikt Brors, Rainer Haas, Roland Eils

Deutsches Krebsforschungszentrum, Im Neuenheimer Feld 280, 69120 Heidelberg, Germany

*Corresponding author Email: f.roels@dkfz.de

A strategy will be outlined that integrates univariate and multivariate testing methods for microarrays with pathway database information. When multivariate testing is done, the gene groups to be tested are usually selected prior to the analysis, either based on pathways that are assumed to be affected or more general characteristics such as kinase activity. Because of this static nature, this approach is unable to go beyond prior assumptions. In the method described here, gene groups are selected based on the data by using gene interaction information and univariate statistics. Univariate testing [2] is performed to get an initial list of genes showing differential expression. The resulting genes are mapped onto a general gene interaction network [1] and used to select network segments. A network segment consist of the mapped gene as a center gene and all the genes it is connected to. The network segments are used for further multivariate testing [3]. The integration with interaction information allows the results to be presented in a biologically meaningful context while being founded statistically.

References

1. Biocarta at nci. http://pid.nci.nih.gov/browse_pathways.shtml#biocarta.

2. R. Breitling, P. Armengaud, A. Amtmann, and P. Herzyk. Rank products: a simple, yet powerful, new method to detect differentially regulated genes in replicated microarray experiments. *FEBS letters*, 573(1-3):83,92, August 2004.

3. JJ. Goeman, SA. van de Geer, F. de Kort, and HC. van Houwelingen. A global test for groups of genes: testing association with a clinical outcome. *Bioinformatics*, 20(1):93,99, January 2004.

Cross platform annotation of sequences with Gene Ontology and pathway terms using Goblet

Detlef Groth*, Stefanie Hartmann, Albert J. Poustka, Georgia Panopoulou and Steffen Hennig

Max Planck Institute of Molecular Plant Physiology, Potsdam, Germany

*Corresponding author Email: dgroth@mpimp-golm.mpg.de

The investigation and annotation of newly sequenced genomes is currently mainly done by using sequence comparisons with genomes of related, well characterized organisms using programs like BLAST [1]. As it is often desired to get a broader overview on the function of genomic sequences terminologies like those of Gene Ontology (GO) [2] or pathway terms are used. Although a variety of online resources exists to simplify the setup and workflow for such sequence annotation with GO or pathway terms, those resources are limited due to server limitations on their high throughput capacity.

Furthermore often the problem occurs that certain resources might be discontinued and subsequent analysis must be performed using other services, making comparisons between different investigations often difficult. We therefore created GOblet which can be used either locally or remotely to annotate sequences with unknown functions [3]. Goblet was programmed in a crossplatform manner using C- and Tcl-code [4].
The system is using the sequences and the GO annotations of the UniProt website [5], as well as the standalone NCBI-BLAST. Mappings between GO and pathway terms are used to integrate pathway annotations as well. Having both a webservice and a standalone version of a tool available makes analysis more reliable and users can continue their analysis even if the webservice might be discontinued or only temporarily out of order. GOblet has been tested and installed on different operating systems like Win32, Linux, OSF-1, Solaris and Mac-OSX. Ports to other platforms can be made easily as long as a modern Gnu-C compiler is available [6]. GOblet, released with an open source license, is freely available at http://goblet.molgen.mpg.de

References:

[1] NCBI-BLAST http://blast.ncbi.nlm.nih.gov/Blast.cgi

[2] Gene Ontology http://www.geneontology.org

[3] GOblet-Webserver http://goblet.molgen.mpg.de

[4] Tcl-programming language http://www.tcl.tk

[5] UniProt-Webseite http://www.uniprot.org

[6] GCC-Compiler [4] http://gcc.gnu.org/

Using data-driven modelling of signal transduction pathways and online pathways databases for prediction of pathways behaviours

Shervin Mohammadi Tari*

School of Computer Science, University College of science, University Of Tehran

*Shervin Mohammadi Tari Email: Shervin.mohamamdi@gmail.com

Abstract

Modelling signal transduction pathway is one of the most significant issues in the systems biology. In this study, I introduced a computer model which is able to use the pathway databases contents by using modelling methods to predict the pathway's fluctuations and behaviours based on the combination of data derived from database and the experimental data. This approach derives the Scientists experimental data about the concentration's pathway species in a snapshot and then combines that information with the specific data about the concentration of the indicator species of the pathway from "Pathway Databases" (such as '*NCICB Pathway Interaction Database*', '*KEGG Pathway Database' and 'SPAD: Signalling PAthway Database*'). Finally using ODEs or stochastic simulation (such as markov chain) over the kinetic formulas of the pathway's inner interactions to predict and figure out the pathway's fluctuations. I consider the problem of modelling a mixture of pathways from number of different pathway species, which can interact through a number of reactions, therefore, I have also used biological Networks to find out other pathways influenced that specific pathway by some known interactions which change the concentration's of specific species depends on the biochemical conditions which consequently change the behaviour of the pathway.

References

[1]E. de Silva, M. P. H. Stumpf. *Complex networks and simple models in biology.* J. R. Soc. Interface (2005).

[2]L. J. Steggles, R. Banks, O. Shaw, A. Wipat. *Qualitatively modelling and analysing genetic regulatory networks: a Petri net approach.* Bioinformatics (2007).

[3]Janes K.A, Yaffe M.B. *Data-driven modelling of signal-transduction networks*. Molecular Cell Biology (2006).

[4]S. Ramsey, D. Orrell, H. Bolouri. *DIZZY: Stochastic simulation of large-scale genetic regulatory networks*. Journal of Bioinformatics and Computational Biology (2005).

[5] Hartwell, L. *Genetics. Robust interactions*. Science (2004).

Estimating the topological role of elements in regulatory networks with the pairwise disconnectivity index

Björn Goeman*, Anatolij P. Potapov and Edgar Wingender

Dept. of Bioinformatics, UMG, University of Göttingen, Göttingen, Germany

*Corresponding author Email: bjoern.goemann@bioinf.med.uni-goettingen.de

While theoretical studies of regulatory networks usually deal with the characterization of their global properties, experimentalists focus on the role of molecules and genes in these systems. To link these apparently opposite views one needs to combine general with individual aspects and convey the abstract topological properties of regulatory systems into testable functional characteristics of their components. Few approaches exist for this purpose but their use is restricted due to their inherent limitations, as in the case of betweenness centrality, and the particular properties of regulatory networks. Here, we introduce a new topological parameter - the pairwise disconnectivity index of a network's element - that overcomes these drawbacks.

The measure quantifies how essential a network element is for sustaining the communication ability between connected pairs of vertices in the global context of a network. Such an element can be a single vertex (i.e. molecule, gene) or edge (i.e. reaction) or a group of vertices or edges. The pairwise disconnectivity index is a centrality of the robustness of a network upon the absence of a considered element, i.e. it tests how sensitive a network is to the removal of an element. We have applied the approach exemplary to the analysis of the transcription regulation networks of E. coli and S. cerevisae, the neuronal connectivity network of C. elegans and TLR4 signaling pathway in mammalian species.

References:

Potapov, A.P., Goemann, B. and Wingender, E.: The pairwise disconnectivity index as a new metric for the topological analysis of regulatory networks. BMC Bioinformatics 9, 227 (2008)

Transcription Factor Binding Site Co-Occurrence as a Means to Predict Transcription Factor Interactions and Regulatory Regions

Holger Klein*, and Martin Vingron

Max Planck Institute for Molecular Genetics
Ihnestr. 63-73, 14195 Berlin

*Corresponding author   Email: holger.klein@molgen.mpg.de

Background

Transcription factors (TF) control transcription by forming complexes that bind to the DNA. Therefore the respective transcription factor binding sites (TFBSs) often occur near to each other on the DNA. In this work we exploit this fact to identify TFs that are likely to interact with each other and to predict regulatory regions in human.

Methods

Identification of Potential Interaction Partners

We annotate known regulatory regions with predicted transcription factor binding sites based on binding site motifs from the TRANSFAC database. We count co-occurring TFBS pairs using a sliding window. Subsequently we define a co-occurrence score S_{ij} as a log-odds score of observed (c_{ij}) and expected ($c_{ij,exp}$) numbers of TFBS pairs. For the expected number of pairs we take the average pair count resulting from a repeated TFBS label permutation procedure.

$$S_{ij} := \log \frac{c_{ij,obs}}{c_{ij,exp}} = \log \frac{c_{ij}}{\sum^{p} c_{ij,perm.}/p}$$

Prediction of Regulatory Regions

Using predicted TFBSs and the co-occurrence scores described above we build TFBS-graphs with binding sites as nodes and the TFBS pair co-occurrence score as the edge weight. We perform graph matching to assess the resemblance of TFBS combinations in the given sequence to the ones in known regulatory DNA and take the resulting sum of edge-weights of the matching as the regulatory potential of a piece of sequence.

Results

We use a non-redundant set of human upstream regions from EnsEMBL and a representative set of vertebrate position weight matrices from TRANSFAC to calculate the co-occurrence scores. The scores for known interactions taken from TRANSFAC get significantly higher scores than combinations not known to interact. Moreover we identify new candidate TF interactions.

We test the parameter space for the co-occurrence procedure and employ the co-occurrence matrix derived from the best parameter combination in the TFBS graph procedure. Furthermore we evaluate different matching algorithms. We show the application of the prediction of regulatory regions for artificial, promoter and enhancer data sets.

Reference

H. Klein, and M. Vingron, Genome Informatics, 18, pp. 109-118, 2007.

Membrane identity and GTPase cascades regulated by toggle and cut-out switches

Perla Del Conte-Zerial, Lutz Brusch*, Jochen C. Rink, Claudio Collinet, Yannis Kalaidzidis, Marino Zerial and Andreas Deutsch

Center for Information Services and High Performance Computing, University of Technology Dresden, 01062 Dresden, Germany

*Corresponding author Email: lutz.brusch@tu-dresden.de

Key cellular functions and developmental processes rely on cascades of GTPases. GTPases of the Rab-family provide a molecular ID-code to the generation, maintenance and transport of intracellular compartments. Here, we addressed the molecular design principles of endocytosis by focusing on the conversion of early endosomes into late endosomes, which entails replacement of Rab5 by Rab7 (Rink et al., 2005). We modelled this process as a cascade of functional modules of interacting Rab GTPases. We demonstrate that inter-module interactions share similarities with the toggle switch described for the cell cycle. However, Rab5-to-Rab7 conversion is rather based on a newly characterised "cut-out switch" analogous to an electrical safety-breaker. Both designs require cooperativity of auto-activation loops when coupled to a large pool of cytoplasmic proteins. Live cell imaging and endosome tracking provide experimental support to the cut-out switch in cargo progression and conversion of endosome identity along the degradative pathway (Del Conte-Zerial et al., 2008). We propose that, by reconciling module performance with progression of activity, the cut-out switch design could underlie the integration of modules in regulatory cascades from a broad range of biological processes.

Del Conte-Zerial, P., Brusch, L., Rink, J.C., Collinet, C., Kalaidzidis, Y., Zerial M. and Deutsch A. (2008) Membrane identity and GTPase cascades regulated by toggle and cut-out switches. Molecular Systems Biology, in press.

Rink, J., Ghigo, E., Kalaidzidis, Y. and Zerial, M. (2005) Rab conversion as a mechanism of progression from early to late endosomes. Cell, 122, 735-749.

A systems biology approach for analyzing RNAi data using functional networks

Angela Simeone and Andreas Beyer**

Biotechnology Center (BIOTEC), Technische Universität Dresden, Tatzberg 47/49, 01307 Dresden, Germany

*Corresponding author Email: angela.simeone@biotec.tu-dresden.de, andreas.beyer@biotec.tu-dresden.de

The improvement and the automation of genome-wide RNAi screens for studying complex cellular processes provides a huge amount of data turning the data analysis and interpretation into a bottleneck.

RNA interference (RNAi) is a method used to specifically suppress gene expression by targeting and degrading mRNA in order to systematically analyse the effects of the loss of function. RNAi screening data are subject to noise because the suppression may be inefficient, because the detection of the phenotype can be inaccurate or due to off-target effects. Moreover, it can be difficult to explain the observed genotype-phenotype association exclusively based on phenotypic data.

In order to address these issues and to correctly understand the role of each gene/protein in the specific (emergent) cellular process we integrate the phenotypic information with independent gene interaction data (functional network) [1,2]. We then screen the network for identifying functional modules associated with respective phenotypic effects.

Results

As a proof of principle we have applied this approach to a recently published RNAi screen, which aimed at detecting cell-cycle related genes in human cell-lines [3].

When applied to the G1-arrest phenotype our method identified a network module of 106 genes. This analysis successfully detected genes that were truly related to G1-arrest as confirmed by independent experiments (true positive). Also, the network module contained genes that did not show a significant phenotype in the primary screen, but which were later confirmed to be also related to G1-arrest (false negatives).

References

[1] von Mering C. et al., Nucleic Acids Research, vol. 35 pp. D358-D362, Jan 2007

[2] Mishra G. et al., Nucleic Acids Research, vol. 34 pp. D411-D414, Jan 2006

[3] Kittler R. et al., Nature Cell Biology, vol.9 pp. 1401-1412, Nov. 2007

Towards an automated platform for researching the homeostasis of epithelial tissue

Thora Pommerencke*, Thomas Sütterlin, Hartmut Dickhaus and Niels Grabe

Institut für med. Biometrie und Informatik, Universität Heidelberg, Heidelberg, Germany

*Corresponding author Email: thora.p@gmx.de

Tissue homeostasis is the highly regulated equilibrium of cell proliferation, differentiation and cell death enabling the constant self renewal of a tissue. Disturbances in this equilibrium can lead to diseases like psoriasis and cancer. For developing effective treatments the profound knowledge of the epithelial tissue homeostasis is essential. Multi-scale systems biological models are ideal for studying such a complex system.

We present first results towards an automated platform for researching the homeostasis of epithelial tissue. This platform links the experiment with an in silico model by image analysis allowing the quantitative measurement of expression patterns in fluorescence stained tissue sections after digitalization by high throughput microscopic scanning. From these patterns spatial and temporal networks are reconstructed underlying epithelial tissue homeostasis or exogenously (e.g. by toxic substances) caused perturbations. The determined networks can then be integrated into a multi-cellular simulation of the skin.

The expression patterns of five differentiation markers described in literature have been measured and used for reconstruction of an initial network of epidermal differentiation. To support the manual specification of regulatory networks we developed a graphical model builder which directly generates source code to be executed in a multi-cellular tissue simulation. The challenge to further combine these two steps into an integrated and automated platform will be conducted in a dedicated Junior Research Group EPISYS in the frame of the BMBF FORSYS PARTNER programme.

Adapted Boolean Network Models for Extracellular Matrix

Johannes Wollbold*, Ulrike Gausmann, Reinhard Guthke, Raimund W. Kinne and René Huber

Institute of Algebra, Technische Universität, Zellescher Weg 12-14, D-01062 Dresden, Germany

*Corresponding author Email: jwollbold@gmx.de

Human rheumatoid arthritis (RA) is characterised by chronic inflammation and destruction of multiple joints. Semi-transformed synovial fibroblasts (SFB) are known to have a decisive influence on development and progression of the disease by predominant expression and secretion of pro-inflammatory cytokines and tissue-degrading enzymes, thus maintaining joint inflammation, degradation of extracellular matrix (ECM) components and invasion of cartilage and bone. In addition, fibrosis of the affected joints is also driven by SFB expressing enhanced amounts of ECM components like collagens.

We applied a new procedure to simulate and analyse the temporal behaviour of regulatory and signalling networks based on formal concept analysis (FCA).[1] We started with literature mining, collecting information about regulatory interrelations of 19 genes closely related to the defined biomedical question. Then we extracted time-course-related information from microarray expression experiments in TNFα- and TGFβ-stimulated SFBs. This was achieved by data discretisation to 0 or 1 using the k-means clustering method.

Simulating the network, we adapted the Boolean functions iteratively to our data according to biologically justified assumptions. The final simulations were further analysed by the attribute exploration algorithm of FCA, integrating again the observed time series in a more fine-grained manner. We obtained two knowledge bases (KB), which contain sets of temporal rules describing cellular responses due to the external stimuli. These KB can be used for further analysis of the ECM system in human fibroblasts and may be queried to predict the functional consequences of observed (e.g. in diseases as RA) or hypothetical (e.g. for therapeutic purposes) gene expression disturbances.

[1] Wollbold J, Guthke R, Ganter B: *Constructing a Knowledge Base for Gene Regulatory Dynamics by Formal Concept Analysis Methods.* In K. Horimoto et al. (Eds.): AB 2008, LNCS 5147. Springer, Heidelberg 2008, pp. 230-244.

Analyzing metabolomic mass spectrometry data with one-dimensional self-organizing maps

Peter Meinicke*, Thomas Lingner, Alexander Kaever, Kirstin Feussner, Cornelia Goebel, Ivo Feussner, Petr Karlovsky, and Burkhard Morgenstern

Georg-August-Universitaet Goettingen, Germany

*Corresponding author Email: peter@gobics.de

One of the goals of global metabolomic analysis is to identify metabolic markers that are hidden within a large background of data originating from high-throughput analytical measurements. Metabolite-based clustering is an unsupervised approach for marker identification based on grouping similar intensity profiles of putative metabolites. A major problem of this approach is that in general there is no prior information about an adequate number of clusters. We present an approach for data mining on metabolite intensity profiles as obtained from mass spectrometry measurements. We propose one-dimensional self-organizing maps for metabolite-based clustering and visualization of marker candidates [1]. In a case study on the wound response of *Arabidopsis thaliana*, based on metabolite profile intensities from eight different experimental conditions, we show how the clustering and visualization capabilities can be used to identify relevant groups of markers. In that context our specialized realization of self-organizing maps is well-suitable to gain insight into complex pattern variation in a large set of metabolite profiles. In comparison to other methods our visualization approach facilitates the identification of interesting groups of metabolites by means of a convenient overview on relevant intensity patterns. In particular, the visualization effectively supports researchers in analyzing many putative clusters when the true number of biologically meaningful groups is unknown.

References

1. P. Meinicke, T. Lingner, A. Kaever, K. Feussner, C. Goebel, I. Feussner, P. Karlovsky, and B. Morgenstern. Metabolite-based clustering and visualization of mass spectrometry data using one-dimensional self-organizing maps. *Algorithms for Molecular Biology*, 3:9, 2008.

Optimization as a framework for metabolic reconstruction

Analysis of suboptimal solutions

Luis F. de Figueiredo[1] *, Francisco J. Planes[2], Stefan Schuster[1]

[1]Bioinformatics Department, Friedrich-Schiller-University Jena, Germany

[2]CEIT and TECNUN, University of Navarra, Spain

*Corresponding author Email: ldpf@minet.uni-jena.de

The modelling and simulation of metabolic networks at various scales is an important field in Systems Biology in the post-genomic era. The metabolic capabilities of small sized networks can be easily elucidated by determining all non-decomposable pathways at pseudo-steady state, either using Elementary Flux Modes [6] or Extreme Pathway [5] frameworks. As the size of the network increases, at best to genome scale, the enumeration of these pathways becomes computationally intractable [4]. In order to deal with this issue, the optimization of a specific reaction flux, typically towards maximal biomass yield, has been proposed [2,3]. An alternative method to analyse metabolism is to detect the minimal subnetwork capable of synthesizing target metabolites given a list of source metabolites and a metabolic network. Beasley and Planes have recently presented a mathematical optimization model that reconstructs such subnetworks, given a source and a sink compound [1]. Herein we extend some of the features of the model and illustrate our refined approach by means of an example from the tricarboxylic acid cycle. We show that the mathematical model presented by Beasley and Planes has a good overall performance and that the results can be improved by changing/adding further constraints. In addition, we analyze glycolysis-like solutions. Our study shows that other pathways stoichiometrically similar to glycolysis exist, with the same ATP yield and less reaction steps.

[1] Beasley and Planes (2007) Bioinformatics, 23:92-98
[2] Edwards et al. (2001) Nature Biotech., 19:125-130
[3] Fell and Small (1986) Biochem. J., 238:781-786
[4] Klamt and Stelling (2002) Mol. Biol. Rep., 29:233-236
[5] Schilling et al. (2000) J. Theor. Biol., 203:229-248
[6] Schuster and Hilgetag (1994) J. Biol. Syst., 2:165-182

Elementary flux modes and optimization

Adam Podhorski[1], Francisco J. Planes[1*], Luis F. de Figuereido[2], Angel Rubio[1], John E. Beasley[3], Stefan Schuster[2]

[1]CEIT and TECNUN, University of Navarra, Spain

[2]Bioinformatics Department, Friedrich-Schiller-University Jena, Germany

[3]Mathematical Sciences, Brunel University, UK

[*]Corresponding author Email: fplanes@tecnun.es

In the post-genomic era elementary flux modes represent a key concept to analyze metabolic networks from a pathway-oriented perspective (Schuster *et al.*, 2000). Despite their early formulation (Schuster and Hilgetag, 1994), the computation of the full set of elementary flux modes in large-sized metabolic networks still constitutes a challenging issue to meet. A summary of the different algorithms proposed to carry out this task can be found in Klamt *et al.*, 2005. Based on the work of Beasley and Planes, 2007, we here illustrate that the full set of elementary flux modes can be enumerated via mixed-integer linear programming. Technically, our approach produces elementary flux modes in increasing number of reactions by sequentially solving an optimization problem. Though our procedure is not particularly efficient for large-sized metabolic networks, it is much more flexible. It can be applied to calculate the elementary flux modes satisfying a given criteria without having to calculate all the solutions first, as typically done by current methods. This greatly speeds up the computations by allowing to focus only on that part of the solution space that is of interest. To illustrate the scope of our approach, we here consider two different cases, namely modes in a given length range and ATP producers modes. Our analysis shows that our mathematical approach can be an effective tool to explore the capabilities of metabolic networks, including those at the genome-scale.

References

Beasley, J.E. and Planes, F.J. (2007) *Bioinformatics,* 23(1), 92-98.

Klamt, S., Gagneur, J. and Von Kamp, A. (2005) *IEE Proc Systems Biol*, 152, 249-55.

Schuster, S., Fell, D.A. and Dandekar, D. (2000) *Nature Biotechnology,* 18, 326-332.

Schuster, S. and Hilgetag, C. (1994) *Journal of Biological Systems* 2, 165-182.

Normalization of Metabolomics Data using Bayesian Networks

Yvonne Pöschl*, Christoph Böttcher, Steffen Neumann, Stefan Posch, and Ivo Grosse

Institute of Computer Science
Martin Luther University Halle–Wittenberg

*Corresponding author Email: poeschl@informatik.uni-halle.de

Metabolomics data are often recruited to investigate the responses of organisms to changes in the environment or to interventions. However, separating the biological variability of the metabolites from the technical variability caused by the analytical platform on which the metabolites are measured is a challenge of modern metabolomics. Here, we propose a normalization method for removing such platform-specific technical variations.

The proposed method is inspired by a publication of Sysi-Aho et al. [1], where a normalization method for metabolomics data is introduced that uses an optimal selection of multiple internal standards (NOMIS). NOMIS uses a fixed number of five predefined internal standards to determine the normalization factor for each mass signal. Here, we extend the approach of Sysi-Aho et al. by allowing (i) the selection of an optimal subset of internal standards for each mass signal separately, and (ii) a fully Bayesian treatment of the resulting optimization problem.

We compare both approachs on the mouse lipidomics data from Sysi-Aho et al., and we apply the extended approach to in-house Ultra Performance Liquid Chromatography coupled to Mass Spectrometry (UPLC/MS) data measured on a Bruker micrOTOF-Q. The in-house dataset consists of measurements of methanolic extracts from *Arabidopsis thaliana* leafs and seeds. Preprocessing, such as peak picking and alignment, was done using the R-package XCMS [2], but the developed method is applicable to peak tables in general.

We find that more than two thirds of the measured mass signals can be normalized by subsets consisting of not more than three internal standards. This indicates that the Bayesian method accomplishes an accurate normalization using substantially fewer parameters.

References

1. M. Sysi-Aho et al. Normalization method for metabolomics data using optimal selection of multiple internal standards. *BMC Bioinformatics*, 8(93), 2007.

2. XCMS 1.12.*, http://www.bioconductor.org.

Integrative Network Reconstruction And Analysis Using TRANSPATH®And KEGG-LIGAND

Ante M.*, Crass T., Goemann B., Wingender E.

Department of Bioinformatics (Medical Faculty)
Georg-August-University Goettingen
Goldschmidtstraße 1
37075 Goettingen

*Corresponding author Email: michael.ante@bioinf.med.uni-goettingen

Biological networks of healthy and diseased tissues may exhibit different topologies. Identifying these differences could render drug development significantly more cost- and time-effective. We will present a proposal for a knowledge-based systems biology approach to the identification of key nodes that are involved in the expression of a given disease phenotype.

To approximate real biological systems we are implementing an integrative network approach that considers how metabolic products can act as signal molecules and signal cascades can lead to changes in metabolism. Therefore instead of focusing on single aspects, we are combining them.

We defined a novel network representation using a bipartite graph. Due to this definition we propose several transformation rules to understand and analyze networks more easily. We used TRANSPATH® [1] and KEGG-LIGAND [2] to populate an integrative database and mapped their entries. We are setting up a pipeline for network reconstruction and analysis. GO [3] and the anatomic ontology CYTOMER [4] are used to filter the networks towards species and tissue specificity. Differential expression data will be used as well as derived information gained from the DEEP tool [5]. Topological analysis will include the recently published pairwise disconnectivity index [6].

Results of comparing healthy and diseased tissue networks will be shown.

References

1. M. Krull et al. *Nucleic Acids Research*, 34:546–551, 2006.

2. S. Goto et al. *Bioinformatics*, 14:591–599, 1998.

3. M. Ashburner et. al. *Nature Genetics*, 25:25–29, 2000.

4. T. Heinemeyer et al. *Nucleic Acids Research*, 27:318–322, 1999.

5. J. Degenhardt et al. *Nucleic Acids Research*, 35:619–624, 2007.

6. A. Potapov et al. *BMC Bioinformatics*, 9:227, 2008.

FOLA: Flux Optimizer via Linear Algebra for Metabolic Simulations

Susanna Bazzani*, Kai Hartmann, Thorsten Fallisch, Dietmar Schomburg

Department of Bioinformatics and Biochemistry, Technical University of Braunschweig
Langer Kamp 19b, 38106 Braunschweig Germany

*Corresponding author Email: s.bazzani@tu-bs.de

1 Abstract

Flux Balance Analysis (FBA) is a well established simulation approach for metabolic networks, based on linear optimization and on the assumption of steady-state [1].

Several software solutions are available for this task [2-4]. However, the customization of these tools is often complex and only few among them are under open-source licence.

Here we present a computational tool that can close this gap.

FOLA is a Linux command line tool, based on the Cubic Metabolome Project (CMP), a C++ library to handle metabolic networks. CMP has been developed in a strict object-oriented perspective and this aspect leads to a robust and easy reusable code.

FOLA is a C++ software, which is capable to detect the dead ends within a network and its FBA simulation. It performs the recursive optimization of all fluxes via Flux Variability Analysis and outputs the results in SBML and MPS (Mathematical Programming System) formats. MPS is a standard ASCII medium for commercial optimization packages. Hence the user can freely decide between lp_ solve or any other preferred software.

FOLA has been tested on *Escherichia coli* core network and on *Streptomyces coelicolor* metabolic network and the obtained growth rates were in agreement with the published data [5-6].

A new version of the CMP library is available on request. The optimization procedures are performed by the open-source solver lp_ solve 5.5.

2 References

1 Feist AM et al., *Mol Syst Biol.* **3**, 121 (2007).
2 Lee D.Y., *Bioinformatics* **19**, 16 (2003).
3 Urbanczik R., *BMC Bioinformatics* **13**, 7 (2006).
4 Klamt S. et al., *BMC Syst Biol.* **8**, 1 (2007).
5 Schilling C.H et al, *Biotechnol Bioeng.* **71**, 4 (2001).
6 Borodina I. et al., *Genome Res.* **15**, 6 (2005).

Numerical analysis of nonlinear, kinetic ordinary differential equation systems for in-silico modeling of biochemical processes

Holger Marquardt, Gabriele Petznick, Paul Hammer, Chong Wang and Peter Beyerlein*

TFH Wildau, University of Applied Science
Bahnhofstr. 1
D-15745 Wildau

*Corresponding author Email: peter.beyerlein@tfh-wildau.de

The study of complex biochemical processes via in-silico modeling has become a promising approach to provide a fundamental understanding of higher biological events. Since this discipline is still young we analyzed an existing model of the human cell cycle developed by Judith Wodke [1].

This model bases mainly on cyclin-dependent kinases (Cdks) as main cell cycle regulators [2] and can be simulated by solving the ordinary differential equation (ODE) system derived from kinetic equations of the system.

A Java framework was developed to simulate and analyze this model. The required information about reaction kinetics were extracted from SBML (Systems Biology Markup Language [3]) and converted into the ODE system as a Java class. The converted ODE system is solved using the Runge-Kutta algorithm by Cash & Karp [4] and the resulting concentration curves are plotted over time.

The original model was able to show the proposed oscillating behavior for about four cell cycle rounds. To enhance this behavior different methods were applied. The substitution of certain compounds with similar artificial curves, which do not depent on the residual ODE system but on time, leads to slightly better behavior. The oscillating behavior was analyzed by Fourier transformation of selected concentration curves. Changes in concentrations between cell cycle rounds were monitored in form of heat maps to determine which compounds begin to vary first and strongest.

Using our analysis results we could improve the oscillating behavior significantly.

References

1. J. Wodke. Qualitative Modelling of the Human Cell Cycle. Master's thesis, Free University Berlin, 2006.

2. A. Csikász-Nagy et al. Analysis of a generic model of eukaryotic cell cycle regulation. *Biophysical Journal BioFAST*, 90:4361–4379, 2006.

3. M. Hucka et al. The systems biology markup language (SBML): a medium for representation and exchange of biochemical network models. *Bioinformatics*, 19:524–531, 2003.

4. J. R. Cash and A. H. Karp. A variable order Runge-Kutta method for initial value problems with rapidly varying right-hand sides. *ACM Transactions on Mathematical Software*, 16:201–222, 1990.

Identification of Rich-Club Components in Eukaryotic Protein Interaction Networks

Beisser, Daniela; Brinck, Heinrich*; Wiebringhaus, Thomas

Institute for Bio- and Chemoinformatics, University of Applied Sciences of Gelsenkirchen, Recklinghausen, Germany

*Corresponding author Email: heinrich.brinck@fh-gelsenkirchen.de

Introduction

Due to the increasing amount of biological data on a molecular level, the analysis of biological networks became an important field to understand the mechanisms of evolution, organisation and function of the cell. Some basic network properties like the scale-free power-law distribution, the small-world behaviour and the hierarchical organisation are also revealed for biological networks.

Aim

For many complex systems, it was recently observed that high degree proteins (hubs) tend to form highly interconnected clusters, so called *rich-clubs*, forming a subgraph of some of the most influential proteins that might provide a backbone allowing a fast communication throughout the network by integrating different functional modules.

Methods

Protein interaction networks of yeast (5294 proteins, 50432 PPIs), fly (7372 proteins, 24922 PPIs) and human (9222 proteins, 36324 PPIs) were compiled from renowned databases.

First quantifications were based on the *rich-club coefficient* as introduced by Zhou and Mondragoran (2003) and subsequently extended by a significance analysis (Colizza et al.). We further extended the measures by ideas coming from Jiang et al. (2008) that we call *rich- club components*.

Results

For all networks, *rich club components* were identified with high statistical significance and biologically examined in detail to analyse unifying functional properties of their members. The largest *rich-club components* for *S. cerevisiae* constitute functions for protein processing, e.g. folding, sorting and degradation. The *rich-club components* of *D. melanogaster* and *H. sapiens* are the MAPK- and the TGF-β pathway, two of the most central and evolutionary ancient pathways for signal transduction that are highly conserved throughout evolution. A high interconnectivity of hubs might cause fast information flow and numerous effective ways of regulation. Moreover, a *rich-club component* with a consistent function might improve overall resilience as neighbouring hubs might compensate malfunctions.

Optimal Probeset Reassembly

C. Hummert[1]*, F. Mech[1], U. Gausmann[2] and R. Guthke[1]

[1]Department Molecular ans Applied Microbiology, Leibniz Institute for Natural Product Research and Infection Biology, Hans-Knoell-Institute, Beutenbergstr. 11a, D-07745 Jena, Germany

[2]Genome Analysis, Leibniz Institute for Age Research, Fritz-Lipmann-Institute (FLI), Beutenbergstr. 11, D-07745 Jena, Germany

*Corresponding author Email: christian.hummert@hki-jena.de

Motivation

The probe design for well established commercial microarrays such as from Affymetrix is often not optimal. This leads to cross-hybridizations. An algorithm to cope this problem has been designed.

Methods

The sequences of all Affymetrix probes are blasted against RefSeq using blastn. For each probe it is counted how many genes matched in the correct orientation from the 5' to the 3' end. All splice variants in the BLAST results were merged for further analysis. All probes matching only one gene are classified as good, and all probes matching more than one gene are classified as bad. Now, only those probes matching exactly one gene are used to build new probesets. The intersection between two probesets is always empty for all probesets. The length of the new probesets is now variable and not fixed to eleven. From these new probesets custom Chip Definition Files (CDF) and the corresponding Bioconductor libraries for Affymetrix human GeneChips are created.

Results

New CDFs have been created for different Affymetrix chips. These CDFs result from Boolean terms, which avoid cross-hybridization completely. They have been compared to the original Affymetrix CDFs and to results from other groups like Ferrari et al. [F07]. The Affymetrix GeneChip data from a concrete experiment from Koczan et al [K08] was analysed using the different CDFs. The obtained results were correlated to qRT-PCR measurement readings for the same experiment. The Boolean probeset Reassembly CDFs have the highest correlation coefficients. Availability: The new CDFs are freely available as R-packages in the Comprehensive R Archive Network (RCRAN) and are easy to use.

References

F07. Ferrari, S. et al.: Novel Definition Files for Human GeneChips Based on GeneAnnot. In: BMC Bioinformatics, 8(1), 446.

K08. Koczan, S. et al.: Molecular Discrimination of Responders and Non-Responders to Anti-TNFa Therapy in Rheumatoid Arthritis. In: Arthritis Research & Therapy, 10, R50.

Improving the Quality of Microarray Analysis with Machine Learning Techniques

C. Hummert[1]*, U. Gausmann[2], R. Guthke[1] and E. G. Schukat-Talamazzini[3]

[1]Department Molecular ans Applied Microbiology, Leibniz Institute for Natural Product Research and Infection Biology, Hans-Knoell-Institute, Beutenbergstr. 11a, D-07745 Jena, Germany
[2]Genome Analysis, Leibniz Institute for Age Research, Fritz-Lipmann-Institute (FLI), Beutenbergstr. 11, D-07745 Jena, Germany

[3]Pattern Recognition, Institute for Computer Science, Friedrich Schiller University of Jena (FSU), Jena, Germany

*Corresponding author Email: christian.hummert@hki-jena.de

Motivation

Microarray experiments always contain high error rates. Systematic errors like cross-hybridizations or intron matching probesets can be overcome by chip design and well done chip definition files (CDFs). On the other hand unsystematic errors introduced by a specific experiment are more difficult to handle [E06]. To cope with this problem an machine learning approach has been developed.

Methods

Data from chip experiments and information about possible error causes are used as input for machine learning algorithms. Feature selection algorithms are planned to be used here. Results from more accurate qtRT-PCR experients are used as teacher signals.
Different approaches like neural networks, support vector machines and statistical learning have been tried as machine learning methods.

Results

The machine learning algorithms are able to generalize the error information calculated from those genes, qtRT-PCR data are available for, to all others. The probesets could be assigned to different categories of reliability. As result untrustable probesets can be ignored and on the other hand the remaining probesets provide more confidence.

References

E06. Michael Eisenstein: Technology Feature Microarrays – Quality Control. In: Nature, 442:31, 1067-1072

Mayday – Powerful and Integrative Microarray Analysis Framework

F. Battke*, S. Symons, P. Bruns, G. Jäger, K. Zimmermann and K. Nieselt

Center for Bioinformatics Tübingen, University of Tübingen, Sand 14, 72076 Tübingen, Germany

*Corresponding author Email: battke@informatik.uni-tuebingen.de

Introduction

DNA Microarrays are the standard method for large scale analyses of gene expression and epigenomics. Analysis software must keep pace with the increasing complexity of generated data. Mayday [1] is a free and flexible graphical workbench for visualization and analysis of microarray data. It is written in Java and can be used as fully-functional WebStart application on every major computing platform without any installation. New challenges can swiftly be met due to Mayday's plugin interface. We recently added several new plugins and improved many core structures for enhanced performance. Currently, Mayday includes a large variety of plugins for visual data exploration, clustering, machine learning (using WEKA [3]) and classification, Gene Set Enrichment Analysis [4] and an interface to the powerful R [6] environment. Mayday can import data from several file formats, database connectivity is included for efficient data organization. Numerous interactive visualization tools, including box plots, profile plots, principal component plots, enhanced heatmap [2], the use of metadata to enhance plots as well as the possibility to create publication quality images make Mayday a power analysis tool for microarray data. Mayday is available at `http://www.zbit.uni-tuebingen.de/pas/mayday/`.

Quality Threshold Clustering

Clustering highlights structures in large data sets. Mayday offers a large choice of partitioning as well as hierarchical clustering methods, such as k-Means, SOM, density-based clustering, Neighbor-Joining, together with many distance measures. We recently added the Quality Threshold clustering method [7]. The method is deterministic and does not require choosing the number of clusters beforehand.

Multiclass Gene Mining

Identifying differentially expressed genes is a task for feature selection algorithms. Mayday's GeneMining plugin offers a range of popular methods including the Quartet Mining [5] algorithm. These methods can now also be applied to multiclass experiments. The resulting gene lists can be used to build classification models.

Extensible Visualization Framework

Visualizations present different views on the same data. Mayday's visualization framework maintains connections between views, allows the inclusion of metadata for more meaningful plots and supports exporting into different pixel- and vector-based file formats. It can easily be extended by new plots.

References

1. J Dietzsch, et al. (2006) Mayday - a microarray data analysis workbench. Bioinformatics 22:1010.
2. N Gehlenborg, et al. (2005) A framework for visualization of microarray data and integrated meta information. Information Visualization 4:164.
3. IH Witten et al. (2005) Data Mining: Practical Machine Learning Tools and Techniques. Morgan Kaufmann.
4. A Subramanian, et al. (2005) Proc Natl Acad Sci U S A 102: 15545.
5. Nieselt-Struwe, K. Graphs in Sequence Spaces: a Review of Statistical Geometry. Biophysical Chemistry 66, 111-131, 1997.
6. R Development Core Team. R: A Language and Environment for Statistical Computing, 2006
7. Heyer, LJ et al. Exploring Expression Data: Identification and Analysis of Coexpressed Genes. Genome Research, 9:1106-1115 (1999).

ResqMi - a versatile software for Resequencing Microarrays

Stephan Symons [a*], Kirstin Weber [b], Michael Bonin [c], and Kay Nieselt [a]

[a] Center for Bioinformatics Tübingen, University of Tübingen, Sand 14, 72076 Tübingen, Germany
[b] Kocherstr. 8, 72768 Reutlingen-Altenburg, Germany
[c] Medical Faculty, Microarray Facility, University of Tübingen, Calwer Str. 7, 72076 Tübingen, Germany

*Corresponding author Email: symons@informatik.uni-tuebingen.de

Introduction

Resequencing microarrays are commonly used for the fast and precise analysis of individual genetic variations. Applications include the identification of genetic diseases by resequencing the respective genes. The analysis of resequencing microarray data, including base calling and evaluation of the called sequence can be time consuming. Especially inspection of uncalled bases (n-calls) is cumbersome, as this requires manual inspection in order not to miss important mutations. Here, we present a new open source software called ResqMi, short for *Reseq*uencing using *Mi*croarrays, which focuses on the efficient and user-friendly analysis, visual inspection and easy manual editing of resequencing microarray data.

ResqMi

ResqMi [3] focuses on resequencing data derived using the Affymetrix GeneChip® Sequence Analysis platform [1]. The user interface of ResqMi was developed focusing on quick navigation and concise overviews. A Resequencing window allows to inspect and edit sequence data, giving a fragment-wise view of the processed data. It also shows a header view which summarizes the reference sequence, and an overview of the called bases highlights n-calls and non-reference calls. The called bases are aligned to the reference sequence, also highlighting non-reference calls. Editing sequences can easily be done in place, just by typing or using a context menu. All other windows are connected with the Resequencing window to adjust to the currently selected position. For swift finding of interesting locations, a bookmark system has been implemented. A tabular showing the calls and quality scores is also available. Reports can be generated from sequence data, that includes an overview of the n calls and the type and position of non-reference calls. Intensity data can be inspected in a tabular view or using lineplots of the currently selected base and its nearest neighbors, allowing to identify possible non-reference calls or regions of low or saturated intensities. For base calling, ResqMi uses a model-based approach as described in [2]. The algorithm calles bases according to a combination of position-wise intensity comparisons as well as a region-wise conformance assessment. Called sequence can be reassessed using Re-Analyze, an algorithm for revising n-calls at positions that allow a clearly distinguishable homozygous call. ResqMi can use mappings of resequencing fragments to genomic or mRNA sequences to give detailed information on the position, including where it is within the gene or if it is polymorphic. ResqMi offers a plugin interface, which allows an easy extension of its functionality. Currently, we have implemented several plugins, including the model-based calling algorithm and Re-Analyze. ResqMi is open source software released under the GPL. Binaries for different platforms and the source code are available at http://www-ps.informatik.uni-tuebingen.de/resqmi.

References

1. Affymetrix, Inc. GeneChip® CustomSeq Resequencing Array Program. Technical report, Affymetrix, Inc, 2004.

2. Richard M. Clark et al. Common sequence polymorphisms shaping genetic diversity in Arabidopsis thaliana. *Science*, 317:338–342, 2007.

3. Stephan Symons et al. Resqmi - a versatile algorithm and software for resequencing microarrays. *Lecture Notes in Informatics*, 2008.

Dynamic Programming Algorithm for Thermodynamically Based Prediction of DNA/DNA Crosshybridisation

S. Torgasin*and K.-H. Zimmermmann

Institute of Computer Technology,
Hamburg University of Technology,
21071 Hamburg, Germany

*Corresponding author Email: torgasin@tuhh.de

Prediction of the secondary structure, that would acquire by annealing two just piecewise complementary DNA strands , is an actual problem of computational structural biology. A careful design of DNA strands is crucial for several biological applications such as microarray techniques, PCR, and DNA computing. For this, the important criterion under laboratory conditions is the hybridization energy of two DNA strands. During the last decade, a thermodynamic model was developed [SantaLucia Jr. *et al.*, 2004] that allows to calculate the DNA/DNA hybridization energy and recently also the crosshybridization energy of structural motifs.

We introduce a new algorithm for the secondary structure prediction of DNA/DNA crosshybridization complexes. The main idea is to use dynamic programming approach defining as subproblems sequence prefixes, ending with complementary bases. Traditionally in same aimed algorithms [Zuker and Stiegler, 1981] the base pair is used as a subproblem. Merging the graph representation of the DNA/DNA complex with traditional sequence grid representation, the problem of finding the complex with lowest energy can be stated as finding minimal path through complementarity dot plot. Such minimal path is a sequence of DNA strutural motifs. The algorithm features :

- considering only the complementary vertexes at the $m \times n$ greed;
- distinct recursive formulas for the vertexes according to their role in possible structure:
    - inside or closing a potential stem,
    - starting a potential stem,

This approach leads to reduction of complexity to $O(n^2)$. Thus the algorithm is suitable to embed into another large-scale applications aiming at design of DNA strands and DNA code.

References

SantaLucia Jr. *et al.*, 2004. SantaLucia Jr.,J. and Hicks,D. (2004) The thermodynamics of DNA structural motifs, *Annu Rev Biophys Biomol Struct*, Vol. 33, pp.415-440.

Zuker and Stiegler, 1981. Zuker, M. and Stiegler, P. (1981) Optimal computer folding of large RNA sequences using thermodynamics and auxiliary information, *Nucleic Acids Research*, **9**, 133-148.

Tyrosine Kinase and Transcription Factor Domains of Physarum and the Origin of Metazoans

Israel Barrantes, Gernot Gloeckner and Wolfgang Marwan*

Max-Planck-Institute for Dynamics of Complex Technical Systems, Magdeburg, Germany

*Corresponding author Email: marwan@mpi-magdeburg.mpg.de

Physarum polycephalum ("slime mold") belongs to the Mycetozoa, a group of amoebozoan protists with high genetic heterogeneity both at the inter- and intraspecies level. Physarum differentiates and shows biochemical features typical of animal cells, like most protein complexes and signaling pathways involved in environmental responses. In order to characterize these protein efectors in Physarum, specifically tyrosine kinases and transcription factors, and to assess the evolutionary relationships between these proteins with their orthologs in other eukaryotes, we carried out a comparative analysis of the protein domain content of P. polycephalum. Using a recently reported transcriptomic set together with the previously available proteomic information of Physarum, we found that some domains related to transcription factors and tyrosine kinases previously only observed in metazoans, can be identified in P. polycephalum, while others appeared to have been acquired by lateral transfer from prokaryotes. Furthermore, in agreement with recent phylogenetic studies, we observed that the phylogenetic position of Mycetozoans may lie closer to the common ancestor of metazoans and Choanoflagellates than Fungi. We expect that the complete picture of the evolution of signal transduction processes in amoebozoa and multicellular organisms may improve as more genomes become available.

RNAMute: Algorithmic Advances and New Applications in RNA Viruses

Alexander Churkin, and Danny Barash*

Department of Computer Science, Ben-Gurion University, Beer-Sheva, Israel

*Corresponding author Email: dbarash@cs.bgu.ac.il

The typical stem-loop secondary structure motif of an RNA virus has been experimentally observed to carry a significant role in both virus replication and translation initiation. For example, in HCV, disruptive mutations were found to cause a structural change that directly leads to either an alternation in virus replication or to a dramatic reduction in translation initiation. Our goal is to computationally predict the minimal number of mutations required to disrupt important secondary structure motifs. Towards this goal, we extend our RNAMute computerized tool such that it is capable of analyzing multiple-point mutations efficiently; we incorporate the genetic code into RNAMute in order to discard less interesting disruptive mutations from the biological perspective; finally, we provide a user-friendly GUI environment for our application. In order to analyze point mutations by their secondary structure, we use Vienna's RNAfold to predict the structure of each mutant and categorize the mutations according to their distance from the wildtype structure. Because employing a brute-force approach in traversing all possible mutations is computationally intensive from the practical standpoint, we propose to first obtain the suboptimal solutions in the secondary structure
prediction run using Vienna's RNAsubopt on the wildtype sequence, and then to stabilize the suboptimal solutions by selective point mutations in the neighborhood of each stem that corresponds to a suboptimal solution observed in the dot plot. This reduces the running time of RNAMute from hours to seconds/minutes. We illustrate our method on the 5BSL3.2 sequence of a subgenomic HCV replicon that was evaluated by site-directed mutagenesis experiments.

Analyzing the Evolution of RNA Secondary Structures in Vertebrate Introns Using Kimura's Model of Compensatory Fitness Interactions

Robert Piskol*and Wolfgang Stephan

Section of Evolutionary Biology
LMU BioCenter
Grosshaderner Str. 2
82152 Planegg-Martinsried
Germany

*Corresponding author   Email: piskol@bio.lmu.de

Previous studies have shown that splicing efficiency, and thus maturation of pre-mRNA, depends on the correct folding of the RNA molecule into a secondary- or higher-order structure [1]. When disrupted by a mutation, aberrant folding may result in a lower splicing efficiency. However, the structure can be restored by a second, compensatory mutation. Here, we present a logistic regression approach to analyze the evolutionary dynamics of RNA secondary structures. We apply our approach to a set of computationally predicted RNA secondary structures in vertebrate introns [4]. Our results are consistent with the hypothesis of a negative influence of the physical distance between pairing nucleotides on the occurrence of covariations [5], as predicted by Kimura's model of compensatory evolution [2]. We also confirm the hypothesis that longer local secondary structure elements (helices) can accommodate a larger number of covariations [3], wobbles and mismatches. Furthermore, we find that wobbles and mismatches are more frequent in the middle of a helix, whereas covariations occur preferentially at the helix ends.

References

1. Y. Chen and W. Stephan. Compensatory evolution of a precursor messenger rna secondary structure in the drosophila melanogaster adh gene. *Proc Natl Acad Sci USA*, 100(20):11499–11504, Sep 2003.

2. M. Kimura. The role of compensatory neutral mutations in molecular evolution. *J. Genet.*, 64:7–19, 1985.

3. J. Parsch, J. M. Braverman, and W. Stephan. Comparative sequence analysis and patterns of covariation in rna secondary structures. *Genetics*, 154(2):909–921, Feb 2000.

4. J. S. Pedersen, G. Bejerano, A. Siepel, K. Rosenbloom, K. Lindblad-Toh, E. S. Lander, J. Kent, W. Miller, and D. Haussler. Identification and classification of conserved rna secondary structures in the human genome. *PLoS Comput Biol*, 2(4):e33, Apr 2006.

5. W. Stephan. The rate of compensatory evolution. *Genetics*, 144(1):419–426, Sep 1996.

IntaRNA: Efficient Target Prediction Incorporating Accessibility of Target Sites

Anke Busch, Andreas S. Richter, and Rolf Backofen*

Bioinformatics Group, Albert-Ludwigs-University Freiburg, Georges-Koehler-Allee 106,
Freiburg, D-79110, Germany

*Corresponding author Email: intarna@informatik.uni-freiburg.de

During the last few years, a multitude of regulatory non-coding RNAs (ncRNAs) have been discovered. Many of these act as post-transcriptional regulators by base pairing to a target mRNA, causing mRNA cleavage or translational repression or activation.

Numerous ncRNAs have already been identified, but the number of experimentally verified targets is considerably fewer. Consequently, computational target prediction is in great demand. Beside the hybridization energy of an RNA-RNA interaction, the accessibility of the target sites has a strong influence on the strength of the interaction. Furthermore, there are several indications that a seed region of perfect complementarity may be required or is at least preferable. Many existing target prediction programs neglect intra-molecular binding and arbitrary seeds, while other approaches are either specialized to certain types of ncRNAs or too slow for genome-wide searches.

We introduce `IntaRNA`, a new general approach to the prediction of RNA-RNA interactions incorporating accessibility of target sites as well as the existence of a user-definable seed. `IntaRNA` has a drastically reduced runtime compared to the best available program for the prediction of general RNA-RNA interactions so far. We successfully applied `IntaRNA` to the prediction of bacterial sRNA targets and determined the exact locations of the interactions with a higher accuracy than competing programs.

DNA Conformations and Their Sequence Preferences

Daniel Svozil*, Jan Kalina,Marek Omelka and Bohdan Schneider

Institute of Chemical Technology Prague, Czech Republik

*Corresponding author Email: daniel.svozil@gmail.com

Abstract

The paper analyzed 3D structures of almost eight thousand DNA dinucleotides from about 500 carefully selected crystal structures, examined possible sequence preferences of various dinucleotide conformers by statistical tests, and investigated how the conformers build larger, often deformed or unusual DNA structures as bent DNA, tetraplexes, and junctions. The analysis identified all the known major conformers of the B, A, and Z types thereby confirming the validity of the procedure but we also observed various intermediate BI/BII and B/A conformers. The BI-form is by far the most populated, BII is a distinct B-form but both BI and BII almost merge in protein complexes by a series of intermediate conformers. In general, proteins significantly broaden the DNA conformation space and induce existence of mixed B/A and even pure A conformers.

The wrapping of DNA around histone proteins in a nucleosome-core particle is attained by a fairly regular alteration of BI and BII conformers, occasionally substituted by deformed BI or B/A conformers. Even in these highly deformed DNA molecules two or more BII conformers only rarely follow each other in sequence.

Some important sequence preferences were observed, here we highlight just a few: Homogenous RR and YY steps (except GG) are over-represented in BI; TG and its Watson–Crick counterpart, CA, prefer BII; the CG and GC steps show a high propensity for mixed B/A conformations. The GC step shows mixed conformational preferences and many GC steps are structurally highly unusual; of all dinucleotides, this step has conformationally the most complicated behavior.

Putative non-coding RNA interactions in brain detected from the porcine EST data

Stefan E Seemann*, Jan Gorodkin

Division of Genetics and Bioinformatics, IBHV, Univ. of Copenhagen, DK-1870 Frederiksberg C, Denmark

*Corresponding author Email: seemann@genome.ku.dk

Proteins make a large network of interactions essential for characterizing their functions and whose deficiency can cause genetic deseases. With the emerging reports of a potential large number of non-coding RNAs (ncRNAs) in the mammalian genome identifying ncRNA components in the networks is of high interest. Here, we aim at identifying such components using our previous work, where we predicted ncRNAs and mRNA like ncRNAs with conserved secondary structures by RNAz in polyadenylated transcribed RNAs in pig [1]. The predictions were based on the Sino-Danish EST resourse which consists of 48,000 contigs (clusters of reads) in 97 non-normalized cDNA libraries in 34 tissues [2]. For any pair of contigs, we searched for patterns of correlated expression between all libraries as well as in the group of 12 libraries comprising only brain and spinal cord. Initially 13,452 protein coding RNAs, 762 putative ncRNAs and 612 combined cis-acting RNAs and ORFs were considered in the search.

Applying hierarchical clustering we found strong correlation between libraries from developmental brain/spinal cord, whereas libraries from adult brain/spinal cord diverge in their expression profile. Coexpressed transcripts were sought by a Hamming like distance measure on expression bitmaps. We detected 29 correlations in which ncRNAs are involved and 805 with combined cis-acting RNAs and ORFs as interaction partners. RNA bindings are tested by their thermodynamic stability.

The putative ncRNAs are weakly expressed and the few ncRNAs with many transcripts are mostly predicted microRNAs. As example a contig with a cluster of two microRNAs is coexpressed with the vehicle-associted membrane protein VAMP2 which is thought to participate in neurotransmitter release. Additionally, we found many known interacting housekeeping genes coding for ribosomal proteins [3] that have RNA structures of low thermodynamic stability.

References

1. S E Seemann et al. *BMC Genomics*, 316(8), 2007.

2. J Gorodkin et al. *Genome Biol*, 8(4):R45, 2007.

3. C von Mering et al. *Nucleic Acids Res.*, 35(Database issue):D358–62, 2007.

Comparative Metagenomics with MEGAN 2.0

Suparna Mitra(1), Daniel C. Richter(1), Alexander F. Auch(1), Stephan C. Schuster(2), and Daniel H. Huson(1)

(1) Center for Bioinformatics (ZBIT), Sand 14, Tübingen, University of Tübingen, Germany
(2) 310 Wartik Laboratories, PennState University, Center for Comparative Genomics, Center for Infectious Disease Dynamics, University Park, PA 1803, USA
Contact: mitra@informatik.uni-tuebingen.de

Metagenomics [1] is a rapidly growing field of research that aims at studying uncultured organisms to understand the true diversity of microbes, their functions, cooperation and evolution, in environments such as soil, water, ancient remains of animals, or the digestive system of animals and humans. A main promise of metagenomics is that it will accelerate drug discovery and biotechnology by providing new genes with novel functions. The recent development of ultra-high throughput sequencing technologies, which do not require cloning or PCR amplification, and can produce huge numbers of DNA reads at an affordable cost, has boosted the number and scope of metagenomic sequencing projects. Increasingly, there is a need for new ways of comparing multiple metagenomics datasets, and for fast and user-friendly implementations of such approaches.

Here we describe some new features of our application MEGAN [2, 3, 4] which has been developed to fit large metagenomic datasets onto a taxonomical tree, thus leading to a visualization of the biodiversity of a given sample. MEGAN provides many options for fine tuning thresholds for high-scoring segment pair selection and taxon matching. These analyses have been enhanced by some new features such as 'rate of discovery', 'functional assessment' and 'meta-data analysis'. In addition to existing features of MEGAN the new version 2.0 allows the comparative analysis of different datasets which can be brought together and compared for taxonomic and functional content. We have developed an interactive and fully customizable chart viewer for MEGAN 2.0 that allows one to extract a number of different comparisons directly from the multiple comparison tree view. A main goal is to provide a tool that can be used to perform large analyses efficiently on a laptop.

[1] Jo Handelsman. Metagenomics: application of genomics to uncultured microorganisms. *Microbiol Mol Biol Rev*, 68(4):669–685, Dec 2004.

[2] Daniel H Huson, Alexander F Auch, Ji Qi, and Stephan C Schuster. Megan analysis of metagenomic data. *Genome Res*, 17(3):377–386, Mar 2007.

[3] Daniel H Huson, Daniel C Richter, Suparna Mitra, Alexander F Auch, and Stephan C Schuster. Methods for Comparative Metagenomics. *Submitted to APBC 2009.*

[4] Hendrik N Poinar, Carsten Schwarz, Ji Qi, Beth Shapiro, Ross D E Macphee, Bernard Buigues, Alexei Tikhonov, Daniel H Huson, Lynn P Tomsho, Alexander Auch, Markus Rampp, Webb Miller, and Stephan C Schuster. Metagenomics to paleogenomics: large-scale sequencing of mammoth DNA. *Science*, 311(5759):392–394, Jan 2006.

Insights into Sequencing Barley - Howto Assemble 454 Reads from Barcoded BAC Pools

Burkhard Steuernagel[1], Andreas Petzold[2], Thomas Schmutzer[1], Stefan Taudien[2], Matthias Platzer[2], Uwe Scholz[1*], and Nils Stein[1]

[1]Leibniz Institute of Plant Genetics and Crop Plant Research (IPK), Corrensstr. 3, 06466 Gatersleben, Germany

[2]Leibniz Institute for Age Research - Fritz Lipmann Institute (FLI), Beutenbergstr. 11, 07745 Jena, Germany

[*]Corresponding author Email: scholz@ipk-gatersleben.de

Measured by acreage barley is the world's top-four crop. Its main uses are for feeding, food production and malting processes. The barley genome with a size over 5,000 mega bases and 80% repetitive DNA is a challenging target for a sequencing project. As an initial start-up we are currently sequencing 3,000 genomic BACs (Bacterial artificial chromosomes) in order to study potential ways to reach the final goal, a sequenced barley genome. The clones chosen for this study fulfill two criteria of being non-overlapping and containing at least one gene.

Using a barcoding strategy and a 454 Genome Sequencer FLX we are able to sequence up to 48 BACs with coverage of at least 15 fold within one run.

For assembling of the raw reads we use MIRA (see http://chevreux.org/projects_mira.html). In our pipeline we mask reads against *E. coli* and vector contaminations. First tests showed that usage of default parameters for the pre-assembly phase in MIRA lead to not optimal results. Furthermore best fitting parameter configuration differs for each clone. Therefore we use a Linux High Performance Cluster for numerical parameter fitting. Results for the first 96 BACs show that we can assemble the raw-data to sets of a few large contigs, in many cases even to a single contig which covers the complete BAC insert. High identity to the reference sequences of five BACs that were obtained with the Sanger method previously, also prove our methods. Currently we are working on a computational evaluation pipeline to rate the assembly results.

Predicting genes on metagenomic pyrosequencing reads with machine learning techniques

K. J. Hoff, M. Tech, and P. Meinicke*

Department for Bioinformatics, Institute for Microbiology and Genetics, University of Göttingen

*Corresponding author Email: peter@gobics.de

Metagenomics provides a set of methods for the genomic characterization of microbial communities. Sequencing of metagenomic DNA yields huge amounts of phylogenetically anonymous DNA fragments with the length of only a few hundred basepairs. A major goal of all metagenomic sequencing projects is the prediction of protein coding genes. Up to now, many metagenomes were sequenced by the chain determination technique with an average read length of 700 bp. Since this technique is costly and limited in throughput, pyrosequencing is gaining increasing popularity for metagenomics. Pyrosequencing currently yields an average read length of 300 bp.

We recently published the successful application of a combination of linear discriminants and a neural network to gene prediction in DNA fragments [1]. In general, the performance of the neural network on 700 bp DNA fragments is comparable to the performance of MetaGene [2], which is another statistical tool for ab initio metagenomic gene prediction. While MetaGene exhibits a higher prediction sensitivity, the neural network shows a higher prediction specificity.

We trained the neural network for pyrosequencing read length and the results indicate that the previously shown trend holds true. Also on 300 bp fragments, the neural network has a higher specificity while MetaGene remains to have a higher sensitivity. Based on these results, we conclude that the neural network can be used for reliable gene prediction on metagenomic pyrosequencing reads.

References:

[1] K. J. Hoff, M. Tech, T. Lingner, R. Daniel, B. Morgenstern, P. Meinicke (2008): Gene prediction in metagenomic fragments: A large scale machine learning approach. BMC Bioinformatics 9:217.

[2] H. Noguchi, J. Park and T. Takagi (2006): MetaGene: prokaryotic gene finding from environmental genome shotgun sequences. NAR 34(19):5623-5630.

Disease Gene Prediction using Phenotypic Similarity based on Semantic Similarity of Gene Ontology terms

Andreas Schlicker*, Thomas Lengauer, and Mario Albrecht

Department of Computational Biology and Applied Algorithmics, Max Planck Institute for Informatics, Campus E1.4, 66123 Saarbrücken, Germany

*Corresponding author E-mail: andreas.schlicker@mpi-inf.mpg.de

Many diseases are caused by mutations in more than one gene. To find such disease genes, linkage and association studies are performed that, however, often yield lists of hundreds of candidate genes that are potentially involved in the disease of interest. Therefore, current computational gene prioritization methods rank those lists of putative disease genes to suggest further validating experiments with the most promising genes at top ranks. In this context, functional similarity has been shown to be one of the most important features for the prioritization method [1]. While Lage *et al.* introduced a text-mining approach for defining a similarity measure of disease phenotypes [2], here we introduce our new medSim score for functionally comparing phenotypes based on the Gene Ontology (GO) annotation of gene products that are known to be associated with the phenotype of interest. The medSim score allows for comparing phenotypes with each other or single proteins with phenotypes. Our results indicate that our method substantially aids the discovery of disease genes and performs quite well in prioritizing candidate genes. To support biological and medical researchers with a simple and fast means of ranking their gene lists, we have extended our FunSimMat web service with an implementation of the medSim score (http://www.funsimmat.de). In general, FunSimMat offers a comprehensive database of precomputed semantic and functional similarity values [3]. This server provides several semantic similarity measures for GO terms as well as diverse functional similarity measures for all proteins from UniProtKB and for all protein families from Pfam and SMART.

[1] Franke L *et al.*, Am J Hum Genet, 2006, 78:1011-1025.
[2] Lage K *et al.*, Nat Biotechnol, 2007, 25:309-316.
[3] Schlicker A and Albrecht M, Nucleic Acids Res, 2008, 36:D434-D439.

GoPubMed: Literature Search with Ontologies

Andreas Doms, Conrad Plake, and Michael Schroeder*

Biotechnological Center of TU Dresden, Tatzberg 47-51, 01307 Dresden, Germany

*Corresponding author Email: ms@biotec.tu-dresden.de

We present GoPubMed [1], a search engine for the biomedical literature that uses the Gene Ontology (GO) and Medical Subject Headings (MeSH) as background knowledge. Both, GO and MeSH, are hierarchical vocabularies for the molecular biology and biomedical domain, respectively. The documents in a search result of a query to PubMed are mapped to concepts in GO and MeSH using text mining techniques. Each time the user selects a concept of interest, only articles related to this concept or its sub-concepts are listed and re-ranked by query keywords and concept names co-occurrences in texts. This sorting and re-ranking of search results allows users to find relevant literature with less effort compared to searching PubMed directly. Since GoPubMed is also constantly analyzing all documents in PubMed, it can present graphs of literature growth over time, as well as key authors, journals, institutions, and cities for each concept/topic in GO and MeSH.

Results

Identification of GO terms in PubMed documents was evaluated on a subset of the GENIA corpus and achieved an F-measure of 85%. A second internal evaluation of 10 use-cases together with biologists revealed that GoPubMed can answer relevant biological questions. GoPubMed is freely available at **www.gopubmed.org**.

References

1. Andreas Doms and Michael Schroeder. GoPubMed: exploring PubMed with the Gene Ontology. *Nucleic Acids Res*, 33(Web Server issue):W783–6, Jul 2005.

Characterization of SNP Variation Effects in Complex Cancer Susceptibility Gene Networks

Erfan Younesi*, Martin Hofmann-Apitius

Bonn-Aachen International Center for Information Technology, University of Bonn, and
Fraunhofer Institute for Algorithms and Scientific Computing (SCAI), Sankt Augustin, Germany

*Corresponding author Email: younesi@informatik.uni-bonn.de

Single Nucleotide Polymorphisms (SNPs) are thought to be the genetic components most responsible for differences in individual disease risk in all major diseases, including breast cancer. Recently the emerging consensus is that combinatorial effects of relevant SNPs may collectively or synergistically contribute to increased breast cancer risk (1). Although this is the assumption underlying most Genome-Wide Association Studies, results of such studies have been mixed and confusing (2).

In an attempt to identify the most important susceptibility players and explore their relationships with other known susceptibility mutations, we used a network-based approach to test the hypothesis that collective topological as well as functional significance of prioritized SNPs might render functional networks (e.g. protein-protein interaction (PPI) networks) unstable and thus might cumulatively promote tumorigenesis.

Reconstruction of human PPI network from validated data repositories using the PIANA software framework (3) allowed us to map breast cancer SNP data from the Cancer Genetic Markers of Susceptibility (CGEMS) initiative (4) onto PPI networks, to generate by this means a susceptibility network, to analyze its topological as well as the functional features, and to incorporate additional information extracted from text mining approaches.

Results

13 susceptibility genes possessing the highest levels of topological and functional significance were identified which showed a cumulative deleterious effect on the stability of a human PPI network. Addition of text mining data suggested novel relationships between these 13 genes and the previously known susceptibility mutations. We can demonstrate that the contribution of collective SNPs to predisposition depends on their topological and functional significance in the network.

References

1. Onay V, *et al*. BMC Cancer. 2006; 6:114.
2. Dimri G. P. Breast Cancer: Basic and Clinical Research 2008:1 1–5.
3. Aragues R, Jaeggi D, Oliva B. Bioinformatics. 2006; 22:1015–1017.
4. http://cgems.cancer.gov/data/

SNP calling by next-generation sequencing on draft genomes

Sebastian Fröhler and Christoph Dieterich*

Max Planck Institute for Developmental Biology, Tübingen, Germany

*Corresponding author Email: chrisd@tuebingen.mpg.de

Introduction

Next-generation sequencing technologies generate millions and more short-length reads in a single run. These data sets facilitate genome-wide analysis of several biological processes. We are interested in whole-genome genotyping of small genomes (~ 10 8 bp), which is, for example, easily performed with a single flow cell on Illumina's Solexa platform. Several computational approaches exist to map and assemble short-length reads onto reference genomes. SNP calling is subsequently performed using the genome maps of the previous step. To our knowledge, none of these approaches considers the quality of the reference genome, which is reasonable for well-studied model organisms like C. elegans and D. melanogaster. Both genome assemblies have undergone "finishing" steps, which involve extensive manual curation. However, the quality of draft genomes may suffer from omitting these finishing steps, which are very costly and therefore often skipped in current pro jects.

Our solution

A simple yet elegant solution is to include the reference genome quality in a common framework for SNP calling. This is most effectively attained in a Bayesian framework, which utilizes sequence quality scores (e.g. Phred scores) from both, short-reads and the reference genome. We decided to build upon a Bayesian framework proposed by Marth et.al. (1999) [1] and extended it to include sequence quality values from the reference genome. This framework calculates the a-posteriori probability of a SNP by integrating information on base background frequencies, base call errors and SNP alignment column permutation probabilities. This approach was further optimized by preprocessing filtering steps.

Results and Conclusions

SNP calling on draft genomes is error-prone if the reference genome quality is ignored. We propose a simple solution to this problem and show its performance on artificial data where we proof our method to perform superior to a frequently used third-party method (MAQ [2]). We also evaluated the performance of our method on the newly sequenced genome of Pristionchus pacificus a satellite organism to Caenorhabditis elegans and confirmed our predictions by traditional Sanger sequencing. Our software is implemented in JAVA and available upon request.

References

1. G. T. Marth, I. Korf, M. D. Yandell, R. T. Yeh, Z. Gu, H. Zakeri, N. O. Stitziel, L. Hillier, P. Y. Kwok, and W. R. Gish. A general approach to single-nucleotide polymorphism discovery. Nat Genet, 23(4):452–456, Dec 1999.

2. Maq: Mapping and assembly with qualities. www.maq.sourceforge.net

Haplotype inference from unphased SNP data in heterozygous polyploids based on SAT

Jost Neigenfind, Gabor Gyetvai, Rico Basekow, Svenja Diehl, Ute Achenbach, Christiane Gebhardt, Joachim Selbig, Birgit Kersten*

GabiPD team, Bioinformatics, Max Planck Institute of Molecular Plant Physiology, 14424 Potsdam, Germany

*Corresponding author Email: kersten@mpimp-golm.mpg.de

Haplotype inference based on unphased SNP markers is an important task in population genetics. Whereas there are different approaches to infer haplotypes in diploid species, the existing software is not suitable to infer haplotypes from

polyploids such as the cultivated potatoes that are tetraploid and highly heterozygous.

Here we present the program SATlotyper [1] which is able to handle polyploid and polyallelic data. SATlotyper formulates “Haplotype Inference by Pure Parsimony” (HIPP) for the Boolean satisfyability problem of propositional logic (SAT). The program was evaluated with simulated as well as experimental SNP data. It was shown that the software is able to infer haplotypes from simulated and experimental input. The experimental data were generated by amplicon sequencing in heterozygous tetraploid populations of potato.

SATlotyper predictions obtained with the simulated data are close to the original simulation. In order to further evaluate SATlotyper, we compared - at the potato locus BA213c14t7 - computed haplotypes comprising 12 SNP positions with experimental haplotypes that were determined by amplicon cloning and sequencing from a sub-population of 19 individuals. 9 of the 10 experimentally derived haplotypes were also computed. With respect to the phased genotypes, the SATlotyper analysis achieved a high correctness compared to the experimental result.

SATlotyper [1] is freeware and available at the Webpage of the GABI Primary Database (GabiPD) at http://www.gabipd.org. SATlotyper will provide haplotype information that can be used in haplotype association mapping studies of polyploid plants.

This work was funded by the German Federal Ministry of Education and Research BMBF (GABI grants 0313112, 0315065B).

[1] Neigenfind J, Gyetvai G, Basekow R, Diehl S, Achenbach U, Gebhardt C, Selbig J, Kersten B (2008) Haplotype inference from unphased SNP data in heterozygous polyploids based on SAT. BMC Genomics, accepted

CNVDB and CNViewer, an integrated system for the analysis and visualization of CNV data.

[1]Jiannis Ragoussis and [1]Ernesto Lowy*

[1]Wellcome Trust Centre for Human Genetics, Roosevelt Drive Oxford OX3 7BN, UK

*Corresponding author Email: ernesto@well.ox.ac.uk

QuantiSNP[1] is an Objective Bayes Hidden-Markov Model to automatically infer regions of copy number variants (CNV) from Illumina's BeadArray^{TM} technology for high-throughput SNP genotyping. Although the importance of these large-scale copy number changes are not fully understood, it is generally accepted that this source of genetic variability may influence phenotype or be involved in disease, affecting gene function (through changing dosage).

For this work, we created CNVDB, a database for CNV data which uses a MySQL backend and CNViewer, a Perl-CGI web interface used to access and display the CNV data in a genomic context.

CNVDB is organised in a hierarchical way where the basic unit of information of the database is the CNV inferred by quantiSNP.

On the other hand, CNViewer, is a tool designed to allow the efficient mining and visualization of CNV data, providing the user with the following capabilities:

- Annotation of CNVs by sample id and copy number change
- Filtering of the predicted CNVs depending on their log Bayes factor value
- Comparison of the CNVs inferred by QuantiSNP with the external CNVs in the latest release of the database of genomic variants (DGV)
- Interactive interface to remove/include individual samples from analysis
- Interactive interface to directly upload, display and analyse (without accessing database) the output files generated by QuantiSNP
- Built-in algorithm to calculate consolidated CNVs loci on-the-fly. Considering a consolidated CNV as a genomic DNA region created by merging overlapping CNVs identified in different individual samples.

Reference:

1. Stefano Colella, Christopher Yau, Jennifer M. Taylor, Ghazala Mirza, Helen Butler, Penny Clouston, Anne S. Bassett, Anneke Seller, Christopher C. Holmes and Jiannis Ragoussis. QuantiSNP: an Objective Bayes Hidden-Markov Model to detect and accurately map copy number variation using SNP genotyping data. Nucleic Acids Research, 2007, Vol. 35, No. 6 2013-2025.

Expression time course analysis of the malaria parasite Plasmodium falciparum

Matthias Scholz* and Martin Fraunholz

Competence Center for Functional Genomics, University of Greifswald, Germany

*Corresponding author Email: matthias.scholz@uni-greifswald.de

The malaria parasite Plasmodium falciparum replicates asexually in an infection cycle within human erythrocytes (red blood cells). By using available time course microarray data we can show that the gene expression data of this intraerythrocytic developmental cycle (IDC) form a circular structure. The trajectory of the data is curved (nonlinear) and closed (circular). To describe the trajectory by a closed curve, we use circular principal component analysis which is a special type of nonlinear PCA (NLPCA). Both methods are nonlinear extensions of standard (linear) PCA, based on auto-associative neural networks.

Circular PCA provides a component which describes the principal curvature of the data by a closed curve. This circular component can then be used as a mathematical model of the developmental cycle of the malaria parasite which gives a continuous and noise-reduced description of the biological process.

The model was used to visualize the transcriptional activity on the parasite's nuclear chromosomes. The analysis revealed a distinct transcriptional activity of telomeric from centromeric genes. This activity delay between genes of chromosome ends (telomeres) and central chromosomal regions suggests that key events of the IDC are initiated at the subtelomeric regions of the P. falciparum chromosomes.

jpHMM: A jumping profile HMM to predict recombinations in HIV-1

Anne-Kathrin Schultz*, Ming Zhang, Ingo Bulla, Thomas Leitner, Carla Kuiken, Bette Korber, Burkhard Morgenstern, Mario Stanke

Department for Bioinformatics, Institute for Microbiology and Genetics, University of Göttingen, Germany
Email: anne@gobics.de

Accurate classification of HIV and the detection of recombinations in HIV genomes are crucial for epidemiological monitoring and developing potential vaccines. Recently we developed jpHMM, a probabilistic model that predicts phylogenetic recombination breakpoints in a query sequence and assigns to each segment of the sequence one HIV-1 subtype. This prediction is based on a multiple sequence alignment of the major HIV-1 subtypes. jpHMM models each subtype in the alignment and allows transitions between different subtypes. The recombination prediction for a sequence is defined by the most probable path through the model that generates the sequence. jpHMM was compared to traditional recombination detection programs and found to perform more accurate for most of the tested sequences.

Previously, our algorithm only output the best guess for the parental subtypes and the recombination pattern. Now, we extended the output to include information on regions where the model is 'uncertain' about the parental subtype and an interval estimate of the breakpoint (breakpoint interval): For each sequence position and each subtype the so-called posterior probability is calculated, that is the probability that this sequence position belongs to the considered subtype given that the whole sequence is generated by the jpHMM. These probabilities are used to define breakpoint intervals and uncertainty regions in the prediction. The user can now be more confident in the predicted parental subtypes outside these regions. For uncertainty regions no parental strain can confidently be determined. However, by examining the graph of the posterior probabilities the user can see which subtypes are most closely related in these regions.

[1] Schultz *et al.*, A Jumping Profile Hidden Markov Model and Applications to Recombination Sites in HIV and HCV Genomes. *BMC Bioinformatics* 7:265. 2006.

[2] Zhang *et al.*, jpHMM at GOBICS: a web server to detect genomic recombinations in HIV-1. *NAR*, 34. 2006.

Classification of HIV-1 Using Coalescent Theory

Ingo Bulla*, Anne-Kathrin Schultz, Fabian Schreiber, Ming Zhang, Thomas Leitner, Bette Korber, Burkhard Morgenstern, Mario Stanke

Department for Bioinformatics, Institute for Microbiology and Genetics, University of Göttingen, Germany
Email: ibulla@uni-goettingen.de

Accurate classification of HIV and other viral sequence data into genotypic subtypes is crucial for epidemiological studies and for developing potential drugs and vaccines. This task is challenging, however, since HIV is one of the most genetically variable organisms known and genomic recombinations are frequent in HIV.

Algorithms for determining whether a particular HIV-1 genome was formed by recombination and,

- if so, detecting the breakpoints and assign the parental strains,
- otherwise, determining the subtype of the examined sequence,

are based on a classification of HIV-1 into genotypic subtypes. Hence, their quality highly depends on the underlying classification. Due to the history of creation of the current HIV-1 nomenclature, it contains several inconsistencies and is somehow arbitrary like all complex classification systems that were created manually. To this end, it is desirable to deduce the classification of HIV systematically by an algorithm.

The main part of our method to obtain such a classification automatically consists in a scoring algorithm which rates a given classification. This algorithm reconstructs ancestral recombination graphs (ARG) of given HIV-1 sequences under restrictions determined by the given classification and finds an ARG with maximal probability by means of Markov Chain Monte Carlo methods. The probability of the most probable ARG is interpreted as a score for the classification. We allow for multiple breakpoints in the recombination events of the ARGs. In the future, we plan to implement an algorithm that improves the classification iteratively with respect to the score.

Our algorithm is applied to the question whether

- subtype G is a pure subtype and CRF02 is a recombinant form having subtype G as one of its parents or
- subtype G is a recombinant form with CRF02 as one of its parental strains and CRF02 is a pure subtype.

Expectation-Maximization - Estimation of Mixture-Densities for the Electron-Spin-Resonance-analysis of Albumin

Carsten Krumbiegel, Andreas Tausche, Paul Hammer, Holger Marquardt, Gabriele Petznick, Chong Wang, Kerstin Schnurr and Peter Beyerlein*

TFH Wildau, University of Applied Science
Bahnhofstr. 1
D-15745 Wildau

*Corresponding author Email: peter.beyerlein@tfh-wildau.de

An early diagnosis of human cancer is of crucial importance for successful therapies. Therefore cancer diagnosis via ESR (Electron-Spin-Resonance) Spectroscopy provides a new promising approach.

A 33-dimensional vector describes a state, i.e. the functional properties of albumin in the blood. This vector is derived from the ESR-Spectrum of a blood sample mixed with so called spin-probes. These spin-probes are bound by albumin. In case of cancer some functions of albumin cause changes in the ESR-Spectrum and thus the vector, which enables differentiation between "healthy" and "suspicious" [1, 2]. It is assumed, that this vector complies with a Gaussian distribution. The respective distribution of "healthy" and "suspicious" patients can be splitted into multiple distributions using the EM (Expectation-Maximization [3,4]) method. This leads to a more accurate for description of the system as well as a decrease of the classification error rate. The focus is on the overlapping area of the "healthy" and "suspicious" distributions, where the classification error rate has its maximum.

This EM method is effected in three different modes: splitting with full covariance, diagonal or identity matrix.

The variations are analyzed with respect to their classification performance and in combination with LDA (Linear-Discriminant-Analysis [5]) to reduce matrix dimensions. The classification compares the maximum values achieved for each Gaussian distribution ("healthy" and "suspicious") where the class with the higher value is chosen.

In addition the analysis is repeated after transformation of normal distributed data to standard normal distributed data. The respective results were compared. Using these methods in combination should decrease the error rate of classification of the two-class-problem ("healthy" and "suspicious").

All methods and corresponding modules where implemented in Perl exclusively.

References

1. S. C. Kazmierczak et al. Electron spin resonance spectroscopy of serum albumin: a novel new test for cancer diagnosis and monitoring. *Clinical Chemistry*, 52(11):2129–2134, 2006.

2. Andreas Tausche et al. Statistical Pattern Classification of Albumine ESR Spectra for Early Diagnosis of Cancer. *German Conference on Bioinformatics*, 2007.

3. Geoffrey J. McLachlan and Thriyambakam Krishnan. *The EM Algorithm and Extensions.* Wiley-Interscience, 1997.

4. Julia Messerklinger. EMAD - EM Clustering mit aggregierten Daten. Master's thesis, Institut für Wirtschaftsinformatik - Data & Knowledge Engineering, 2006.

5. Hans Rudolf Schwarz and Norbert Köckler. *Numerische Mathematik.* Teubner, 2004.

Discovery of biomarkers in potatoes by unbiased variable selection approach

Matthias Steinfath*[1], Nadine Strehmel[2], Joachim Kopka[2], Peter Geigenberger[3], Joost van Dongen[2], and Joachim Selbig[1]

Universität Potsdam, Institut für Biochemie und Biologie, Am Mühlenberg 1, 14476 Potsdam-Golm[1], Max-Planck-Institut für molekulare Pflanzenphysiologie, Am Mühlenberg 1, 14476 Potsdam-Golm,[2],Institut für Gemüse- und Zierpflanzenbau Großbeeren & Erfurt e.V. (IGZ) Theodor-Echtermeyer-Weg 1 14979 Großbeeren[3]

*Corresponding author Email: steinfath@mpimp-golm.mpg.de

Molecular biomarkers are substances used to predict phenotypic traits. They are used in medicine for diagnostic purposes, but also for plant breeding programs.

These makers can be genes, proteins or metabolites. One of the most challenging tasks is to find such molecules, which are predictive for both a large set of genotypes and different environmental conditions.

We present here a procedure allowing us to find markers, which are predictive for potato properties for large set of cultivars and different soils and climate conditions. Metabolite, protein, and transcript profiles were used as input data.

We compared different variable selection and statistical learning methods. We focused on partial least squares (PLS) and variables importance in the projection (VIP) method, which is an embedded variable selection method. The PLS is identifying the combination of the original variables with highest covariance with the trait under investigation. The VIP of each variable considers both correlation with the traits as well as variance within the molecular data.

The variable selection and statistical learning methods were applied to a set of cultivars to each environmental condition separately. Thus, the generality of the resulting models and detected markers could be validated.

The results were also compared to biochemical knowledge from former works. For systems biological interpretation of the results, small models were constructed based on the data.

Genome-wide characterization of human mutually exclusive exons

Martin Pohl*, Dirk Holste, Ralf Bortfeldt, Konrad Grützmann, Stefan Schuster

Friedrich-Schiller-University, chair of Bioinformatics Prof. S. Schuster

* Corresponding author, Email: map@minet.uni-jena.de

Alternative splicing plays an important role in protein diversification in higher eukaryotes. It accounts for different functions and localizations of transcripts from one single gene [1]. Furthermore, it is involved in regulation of gene expression via nonsense-mediated RNA decay. Mis-splicing could be related to important diseases [2]. Alternative splicing that yields mutually exclusive exons (MXE) has not yet been intensively investigated on a genome wide range.
Our analysis is based on alignments of mRNA sequences and expressed sequence tags to the human genome from the UCSC database [3].

Results

We found more than 1300 pairs of MXE, which affect about 4% of the considered genes. Surprisingly, for only 14 of these cases the involved exons lie adjacent. The lengths and distances of the exon pairs show a wide variety, suggesting a low selective pressure and/or the passage of a long period of time since their origin. Adjacent MXE remarkably often encode for subunits of transmembrane molecule transporters and receptor coactivators. MXE of higher order, which involve more than two exons, could not be found. This supports the suggestion of Graveley et al. [4] that vertebrates in general do not have this feature.

References

[1] Lareau et al. The evolving roles of alternative splicing. Curr. Op. in Struct. Biol., 2004, 14:273–282

[2] Ferreira et al. Alternative splicing: a bioinformatics perspective. Mol. BioSyst., 2007, 3:473–477

[3] Karolchik. et al. The UCSC Genome Browser Database. Nucl. Acids Res., 2003, 31(1):51-54.

[4] Graveley et al. Mutually Exclusive Splicing of the Insect Dscam Pre-mRNA Directed by Competing Intronic RNA Secondary Structures. Cell. 2005, 123(1): 65–73.

BACOLAP: fast BAC overlap alignment based on bit-vectors

Krause, J.-U., Kleffe, J.*

Institute for Molecularbiology and Bioinformatics, Charité, Berlin

*Corresponding author Email: juergen.kleffe@charite.de

Introduction

Genome Projects using clone by clone sequencing assemble BACs to form super BACs. But the programs commonly used for short reads can not handle BAC size sequences ranging from 100 to 300 kb. The well-known fast heuristic local alignment tools often miss to find complete sequence overlaps. We therefore developed the program BACOLAP that directly searches for overlapping regions of two BACs with bounded relative edit distance. Based on bit-vector technique the program is nearly as fast as heuristic local alignment software like BL2SEQ [Tat99], NUCmer [Del02], YASS [Noe05] and ClustDB [Kle07] and as accurate as exact global and semi global alignment such as GAP3 [Hua03] and WABA [Ken00]. A post processing of derived overlap alignments helps to identify poorly matching regions which are important to know for the identification of sequencing errors and misassembled BACs.

Methods

Assuming BAC B extends BAC A our new program BACOLAP uses a combination of Myers [Mye99] linear time bit-vector algorithm and Ukkonen's [Ukk85] cut-off heuristic for approximate string matching to compare the first m characters of sequence B with all of sequence A using edit distance in order to detect potential start positions of the overlap in sequence A. Next, using the same method, the last m characters of sequence A are compared with a limited region of sequence B in order to identify possible ends of sequence overlaps. In both cases the edit distance may not be greater than e. The assumed maximal error rate limits the length difference of the two overlapping sequence sections. Finally a combination of Myers [Mye99] bit-vector algorithm and the divide and conquer method by Hirschberg [Hir75] quickly calculates the exact edit distance alignment for given start and end positions in sequences A and B, respectively. The latter algorithm was developed by Aiche et al. [Aic07]. Not seldom the global alignment function produces poor matching regions with lots of randomly distributed gaps in one of the sequences. Such regions are often caused by different repeat copy numbers or inserts and we developed a special post-processing for their identification.

Results

Using a PC with a Pentium 4 processor and 2GB of memory running under Linux OS we compared BACOLAP with ClustDB [Kle07], BL2SEQ [Tat99], WABA (Wobble Aware Bulk Aligner) [Ken00], YASS [Noe05] and NUCmer [Del02] using parameters $m = 300$ and $e = 100$. GAP3 [Hua03] took more than 30 minutes for a single alignment and was not considered for this reason. The parameters for the other programs were gap cost $= -2$, gap extension cost $= -1$, mismatch cost $= -1$ and match cost $= 1$ for BL2SEQ, default parameters for WABA and YASS, $minmatch = 300$ and $maxgap = 100$ for NUCmer, and 100 errors within each alignment window of size 300 for ClustDB. The test set consisted of 1184 BAC pairs from the medicago sequencing project (www.medicago.org) that were supposed to overlap. Table 1 shows how many overlaps the different programs found and how much time they required for computation. BACOLAP identified the largest number of overlapping BAC pairs. Relative to BACOLAP we also list the number of cases in which other programs found the same results or results differing by not more than 20 or 100 sequence positions, respectively. The results agree for most cases but time of computation differs much. The programs YASS and ClustDB are closest to BACOLAPS's sensitivity but take more than twice the time. WABA is clearly the slowest program but more sensitive than NUCmer and BL2SEQ.

Table 1: Comparison between BACOLAP, YASS, ClustDB, WABA, NUCmer and BL2SEQ.

	BACOLAP	YASS	ClustDB	WABA	NUCmer	BL2SEQ
confirmed overlaps	1165	1138	1133	1123	1119	1114
results compared with BACOLAP						
equal positions		1103	1093	1086	1081	1059
position difference maximal 20		34	36	35	37	45
position difference between 21 and 100		1	4	2	1	10
unconfirmed overlaps	19	46	51	61	65	70
computation time	16 min	44 min	42 min	43h 59 min	11 min	12 min

Discussion

We have shown that BACOLAP is nearly as fast as heuristic local alignment programs while identifying more overlaps. We therefore believe BACOLAP to be an useful diagnostic tool for BAC-by-BAC genome assembly projects, which test large numbers of indexed BAC pairs for the overlapping property.

References

Aic07. Aiche, S., Döring, A., Kleffe, J. Fast And Exact Global Sequence Alignment. German Conference on Bioinformatics (GCB2007), September 2007. Potsdam, Germany.

Del02. Delcher, A.L., Phillippy, A., Carlton, J., Salzberg, S.L. Fast algorithms for large-scale genome alignment and comparison. Nucleic Acid Research, 30:2478-2483, 2002.

Hir75. Hirschberg, D.S. A linear space algorithm for computing maximal common subsequences. Commun. Assoc. Comput. Mach., 18(6):341-343, June 1975.

Hua03. Huang, X., Chao, K.-M. A generalized global alignment algorithm. Bioinformatics, 19(2):228-233, 2003.

Ken00. Kent, W.J., Zahler, A.M. Conservation, regulation, synteny, and introns in a large-scale c.briggsae-c.elegans genomic alignment. Genome Research, 10:1115-1125, 2000.

Kle07. Kleffe, J., Möller, F., Wittig, B. Simultaneous identification of long similar substrings in large sets of sequences. BMC Bioinformatics, 8, 2007. (Suppl.5):S7.

Mye99. Myers, G. A fast bit-vector algorithm for approximate string matching based on dynamic programming. Journal of the Association for Computing Machinery, 46(3), Mai 1999.

Noe05. Noe, L., Kucherov, G. Yass: enhancing the sensitivity of dna similarity search. Nucleic Acid Research, 33:540-543, 2005.

Ukk85. Ukkonen, E. 1985. Algorithms for approximate string matching. *Information and Control 64 (1985)*. pp. 100-118.

Tat99. Tatusova, A.T., Madden, T.L. Blast 2 sequences - a new tool for comparing protein and nucleotide sequences. FEMS Microbiol Lett., 174:247-250, 1999.

SynBlast: assisting the analysis of conserved synteny information

Jörg Lehmann*, Peter F. Stadler, and Sonja J. Prohaska

Bioinformatics Group, Department of Computer Science, University of Leipzig, Härtelstraße 16-18, D-04107 Leipzig, Germany

*Corresponding author Email: joe@bioinf.uni-leipzig.de

Motivation: Within the last years more than 20 vertebrate genomes have been sequenced, and the rate at which genomic DNA information becomes available is rapidly accelerating. One of the key tasks in comparative genomics is the retrieval of evolutionarily related genes within and between those genomes. The distinction of orthologous and paralogous sequences, that result from speciation and duplication events, respectively, is important for downstream analysis. Gene duplication and gene loss events inherently limit the accuracy of ortholog detection based on sequence similarity alone. Fully automated methods for ortholog annotation do exist but often fail to identify individual members in cases of large gene families, or to distinguish missing data from traceable gene losses. This situation can be improved in many cases by including conserved synteny information.

Results: Here we present the `SynBlast` pipeline [1] that is designed to extract and evaluate local synteny information. `SynBlast` uses the genomic region around a focal reference gene to retrieve candidates for homologous regions from a collection of target genomes via similarity searches (`BLAST`) and gene content comparisons to the reference region. The candidate regions are evaluated and ranked based on the available evidence for homology including conserved synteny information. The pipeline is intended as a tool to aid high quality manual annotation in particular in those cases where automatic procedures fail. We demonstrate how `SynBlast` is applied to retrieve orthologous and paralogous clusters using the vertebrate *Hox* and *ParaHox* clusters as examples.

Software: The `SynBlast` package written in `Perl` is available under the GNU GPL at http://www.bioinf.uni-leipzig.de/Software/SynBlast/.

References

1. Lehmann J, Stadler PF, and Prohaska SJ. SynBlast: assisting the analysis of conserved synteny information. *BMC Bioinformatics*. Accepted.

Prediction of operons with support vector machines

Christin Weinberg[1], Matthias Scholz[1], Volkmar Liebscher[2] and Martin Fraunholz[1]
Ernst–Moritz–Arndt–University, F.–L.–Jahnstr. 15/15a, D –17487 Greifswald, Germany
[1]Competence Centre for Functional Genomics (CC–FG),
[2]Institute for Mathematics and Informatics

Abstract. Operons are comprised of co–transcribed genes within bacterial chromosomes. Transcription of operons is driven by a promoter and results in a primary transcript that encodes for one or more genes (multi–cistronic operons). As genes with related functions are often organized within operons, the prediction of operons will assist in the functional classification of genes and help to elucidate mechanisms of gene regulation. Since large proportion of the coding sequences (CDS) of newly genomes are ‘hypothetical proteins‘ with no known homolog in the available nucleotide databases – as more genomes become available efficient functional annotation of these genomes will be critical in interpreting the data. Most prediction algorithms applied during annotation and analysis of genome sequences rely solely on indentifying CDS.

We make use of a support vector machine (SVM, [1]) for the prediction of operons within *Staphylococcus aureus COL* [2] a pathogenic bacterium. SVMs are efficient algorithms used in pattern recognition and data classification. The discrimination process is characterized by the maximal margin classifier. As input vectors we employ various genome features like intergenic distance, codon adaptation index, etc.. Additionally we used experimental Northern Blot data of published operon data sets as labels.

References:

1. Burges, Christopher J.C., A Tutorial on Support Vector Machines for Pattern Recognition, 1998, Data Mining and Knowledge Discovery, **2**(2):121–167

2. K. Rogasch, V. Rühmling, J. Pané–Farré, D. Höper, C. Weinberg, S. Fuchs, M. Schmudde, B. M. Bröker, C. Wolz, M. Hecker, and S. Engelmann, *The influence of the two component system SaeRS on global gene expression in two different Staphylococcus aureus strains*. J BACTERIOL 2006, **188**:7742-7758

Array-based prediction of DNA copy number variants for Arabidopsis ecotypes using Hidden Markov Models

Michael Seifert[1], Ali Banaei[1], Jens Keilwagen[1], François Roudier [2], Vincent Colot[2], Florian Michael Mette[1], Andreas Houben[1], Ivo Grosse[3], and Marc Strickert[1]

[1]Leibniz Institute of Plant Genetics and Crop Plant Research, Gatersleben, Germany
[2]Ecole Normale Supérieure, Paris, France
[3]Martin Luther University, Institute of Computer Science, Halle, Germany

Contact: seifert@ipk-gatersleben.de

Arabidopsis thaliana is a model organism in plant biology with a broad geographic distribution including ecotypes from Europe, Asia, and Africa. The natural variation of different ecotypes is expected to be reflected to a substantial degree in their genome sequences. Array comparative genomic hybridization (Array-CGH) can be used to quantify the natural variation of different ecotypes at the DNA level for providing basics to establish a genome-wide map of DNA copy number variants of different ecotypes. Here, we present a new approach based on Hidden Markov Models (HMMs) to predict copy number variants of such Array-CGH experiments. Using the HMM approach, an improved genome-wide characterization of DNA segments with decreased or increased copy numbers is obtained in comparison to the routinely used SegMNT algorithm [1]. Additionally, we investigate which DNA regions are mostly affected by DNA copy number variants exploiting the TAIR8 genome annotation.

[1] Roch NimbleGen, Inc. (2008). A Performance Comparison of Two CGH Segmentation Analysis Algorithms: DNACopy and segMNT. (http://www.nimblegen.com)

Classification of 3-base genes' periodicity reveals traces of reading frame's shifts

Frenkel FE*, Korotkov EV

Bioengineering Centre of RAS, Moscow, Russia

*Corresponding author Email: felix.frenkel@gmail.com

A new concept of triplet periodicity class (TPC) and a measure of similarity between such classes were introduced. We performed classification of 472 288 triplet periodicity (TP) regions found in 578 868 genes from KEGG databank. Totally 2 520 classes were obtained. They contain 94% of 472 288 found cases of TP i.e. major fraction of TP found in genes can be reduced to two and a half thousand classes. For 92% of TP regions contained in classes the same linkage of TP to reading frame (RF) is observed. Each class on the average contained 86.2% of matrices with first base corresponded to first base of codon, and only few classes (about 2%) contained less than 50% of matrices without shift. Therefore there exists strong correlation between TP and ORF in gene. If such correlation does not exist, then formation of TPCs by the algorithm used will be impossible. For 8% of TP cases we revealed a shift between RF of a gene and RF common for majority of genes contained in a TPC. For these 8% of periodic regions the hypothetical amino acid sequences corresponding to RF built by TPC were made. BLAST program has shown that 2 679 hypothetical amino acid sequences have statistically significant similarity with proteins from UniProt databank. We suppose that 8% of TP regions contained in classes possess a mutation originating from RF shift. Such shift or inversion could be a consequence of nucleotide deletions and insertions or DNA sequence inversions when new RF and new amino acid sequence arose. Obtained TPCs can also be used for identification of genes' coding regions and for searching for mutations arisen from RF shift.

References

1. Frenkel FE, Korotkov EV. Classification analysis of triplet periodicity in protein-coding regions of genes. Gene, 2008. Vol. 421(1-2), pp. 52-60. http://dx.doi.org/10.1016/j.gene.2008.06.012.

Maximum conditional likelihood decomposition

Jan Grau*, Diana Boronczyk, Jens Keilwagen, Stefan Posch and Ivo Grosse

Institute of Computer Science
Martin Luther University Halle–Wittenberg

*Corresponding author Email: grau@informatik.uni-halle.de

With the goal of improving the computational recognition of splice sites, we propose a combination of the decision tree model of the maximal dependence decomposition (MDD) algorithm [1] with a discriminative learning approach, resulting in the maximum conditional likelihood decomposition (MCLD) algorithm. The MCLD algorithm learns the structure of the binary decision tree as well as the parameters of the position weight matrix (PWM) models at the leaves in a discriminative manner, maximizing the conditional likelihood as objective function. To obtain a concave objective function, we advance an alternative parameterization of graphical models [2] for the decision tree model.

We compare the accuracy of the MCLD algorithm to popular algorithms for splice site prediction including weight array matrix models, permuted variable length Markov models, maximum entropy models, and the original MDD algorithm on an annotated dataset of mammalian donor splice sites [3], and we find that the MCLD algorithm yields the highest classification accuracy, regarding false positive rate, positive predictive value, area under the ROC curve, and area under the precision-recall curve. Interestingly, the MCLD algorithm accomplishes a more accurate prediction of donor sites than the MDD algorithm with almost tenfold smaller decision trees, consequently resulting in a drastically lower number of parameters. Scrutinizing the MCLD trees, we find several dependencies that are not detected by the original MDD algorithm.

References

1. Christopher Burge and Samuel Karlin. Prediction of complete gene structures in human genomic DNA. *Journal of Molecular Biology*, 268(1):78–94, 1997.

2. Hannes Wettig et al. On supervised learning of bayesian network parameters. Technical Report HIIT 2002-1, Helsinki Institute for Information Technology, 2002.

3. Gene Yeo and Christopher Burge. Maximum Entropy Modeling of Short Sequence Motifs with Applications to RNA Splicing Signals. *Journal of Computational Biology*, 11(2-3):377–394, 2004.

The evolution of the endocytic machinery – a phylogenetic perspective using PYPer

Sabine Bernauer[1], Robert Henschel[2], Matthias Mueller[1], Marino Zerial[1] and Bianca Habermann[1,3*]

1 MPI of Molecular Cell Biology and Genetics, Pfotenhauerstrasse 108, 01307 Dresden
2 Center for Information Services and High Performance Computing (ZIH), 01062 Dresden
3 Scionics Computer Innovation, Tatzberg 47-52, 01307 Dresden

*Corresponding author Email: habermann@mpi-cbg.de

The reconstruction of the evolutionary history of large protein families is crucial for the analysis of homology and inference of function in complex biological systems. With the increasing availability of a large number of complete genome sequences, we now have the possibility to look at the evolution of biological systems in a multitude of ancient and modern organisms. In order to complete this task in a meaningful timeframe, it is however required to streamline and optimize the process of sequence retrieval, alignment and phylogenetic analysis using different resources and bioinformatics tools.

In order to reconstruct the evolution of the endocytic machinery, we have developed PYPer, a generic web-based tool for integrated semi-automated phylogenetic analysis. PYPer performs homology based sequence searching and retrieval, the generation of multiple sequence alignments, automated alignment editing and the generation of phylogenetic trees. Orthologous and homologous sequences are determined by reciprocal Blast and PSI-Blast searches, respectively. Multiple sequence alignments of the related sequences is done using Clustalw and MPI-align (an novel in-house algorithm for the generation of multiple sequence alignments based on Blast pairwise alignments). These alignments provide the basis for phylogenetic construction using various maximum likelihood, parsimony and distance based methods from the Phylip package. In order to provide high flexibility, the user is able to sequentially follow the steps involved in reconstructing the phylogenetic history of a protein family and influence the workflow at any stage interactively.

We have tested PYPer using a variety of endocytic proteins and will present results on Rab GTPases, all members of the VPS9 family, which are nucleotide exchange factors for Rab5, as well as the Rab5 and Rab4 effectors Appl, Rabenosyn and Rabex5.

Molecular systematics in the genomic age: getting mixed signals

Stefanie Hartmann*, Ralph Tiedemann, Joachim Selbig, Christoph Bleidorn

Department of Bioinformatics,
Institute of Biochemistry and Biology,
University of Potsdam
Karl-Liebknecht-Str. 24-25,
14476 Potsdam, Germany

*Corresponding author Email: stefanie.hartmann@uni-potsdam.de

Introduction

In molecular systematics analyses, multiple sequence alignments of homologous sequences are used to infer evolutionary relationships of the organisms they represent. Traditionally, single- or low copy number genes have been used for these studies. It has become clear, however, that a gene tree does not necessarily correspond to a species tree, and that gene trees often disagree with one another.

Factors that contribute to these apparently conflicting signals for single-gene phylogenies (or phylogenies relying on a few genes) include violation of the orthology assumptions, biases leading to non-phylogenetic signal, and stochastic error related to gene length.

Contrary to initial hopes, it has become clear that increasing the amount of sequence information analyzed (i.e., using phylogenomics approaches) does not overcome problems of conflicting signal. Instead of resolving the tree of life, phylogenomic analyses often reveal highly conflicting signals within data sets.

Methods

To aid in determining the cause of conflicting data in phylogenomic data sets, we systematically analyzed gene families by correlating their phylogenetic signal with their functional class. We pay special attention to the practical problems of phylogenomics approaches when the focal taxa are non-model systems, i.e., for which only limited sequence data is available. We apply our framework to the contentious phylogenetic position of Myzostomida, a group of marine invertebrates.

Results and Discussion

Consistent with previous analyses, our combined analysis of almost 40 gene trees using concatenated alignments was unable to resolve the phylogenetic position of Myzostomida. To investigate the source of conflicting phylogenetic signal regarding this taxon, we evaluated topologies of approximately 130 individual gene trees, and we correlated these with evolutionary rates and functional annotation of the genes.

Our results demonstrate that the phylogenetic signal of the single gene tree analyses are generally not a result of artifactual clustering due to long branch attraction. Furthermore, the enrichment analysis of GO IDs supporting a given phylogenetic hypothesis was consistent with previously established developmental and ultrastructural data.

Novel Glocal Alignment of protein sequences with evolutionarily conserved subsequences.

Sriram Balasubramanian*

Indian Institute of Technology, Kanpur, India

*Corresponding author Email: sriramb@iitk.ac.in

A novel Global Local Alignment Algorithm addressing the problem of aligning two protein sequences globally while conserving evolutionarily unchanging, smaller subsequences occurring in one of the sequence, is proposed. Within these subsequences, insertions and deletions are restricted implying evolutionary conservation of structure. Thus the alignment is of global character with local restrictions about these constrained subsequences and hence glocal. This problem is of importance to protein threading applications.
The algorithm is implemented with the methods dynamic programming and iteration. The Needleman Wunsch Algorithm combined with greedy algorithms is extensively used and the Glocal algorithm was implemented successfully on small sequences. It is observed that the primary version of Glocal algorithm, ALGX, works best in the limit that the conserved sequences are smaller in size (approx 1/10th the sequence size) and in number.
A biological interpretation with an evolutionary basis, of the logic of the algorithm is provided. This includes a step by step analysis of the algorithm from a biological and physical point of view of protein alignment.
An intensive theoretical time complexity analysis of the algorithm is also performed.

References:
Needleman & Wunsch (1970) J of Mol Bio
Russel Doolittle (1981) Science
Smith & Waterman (1981) J of Mol Bio
Pearson & Lipman (1985) Science
Pearson & Lipman (1988) PNAS
Altschul et. (1990) J of Mol Bio
Altschul et. (1997) Nucleic Acids Research
Stadin (1997) Nucleic Acids Research
Ralf Thiele, Zimmer (1999) J of Mol Bio
Anna Pachenko et.(1999) Proteins
Jiye , Blundell, Kenji (2001) J of Mol Bio

Wisdom of Crowds: Sum of Kernel Ranks Improves Detection of Remote Protein Homologs

Marten Jäger, Sebastian Bauer, Sebastian Köhler, Peter N. Robinson*

Institute for Medical Genetics, Charité Universitätsmedizin Berlin, Augustenburger Platz 1, 13353 Berlin, Germany

*Corresponding author Email: peter.robinson@charite.de

Protein homology allows one to predict the functional and structural properties of novel proteins by comparison with known ones and therefore plays a vital role in bioinformatics. Discriminative methods such as Support Vector Machines (SVMs) using string and local-alignment (Smith-Waterman) kernels have been shown to be effective for protein homology prediction. Averaging many independent and diverse opinions (”Wisdom of Crowds”) can outperform even expert opinion, as the Victorian polymath Francis Galton first observed a century ago, after the crowd at a county fair accurately guessed the weight of an ox when their individual guesses were averaged.

In this work, we introduce methods to combine the ranks from multiple kernels based on order statistics or simple averages. We tested this method on a comprehensive set of 170 protein families from SCOP. The best individual spectrum kernel ($k = 3$) had a mean ROC score of 0.82 and mean ROC$_{50}$ of 0.38. In contrast, the mean ROC score obtained by combining ranks from nine spectrum kernels for $k = 2, \ldots, 10$ was 0.97, and the mean ROC$_{50}$ score 0.75. Similar improvements were found for combining multiple (k, m)-mismatch kernels as well as multiple Smith-Waterman kernels that had been calculated using different parameters.

Therefore, we conclude that the ’Wisdom of Crowds’ is suitable for homology detection. The aggregation of various SVM methods often results in a higher prediction accuracy than these of a single classifier in the superfamily recognition in SCOP. Furthermore, we showed that the combinations could be calculated very efficiently in linear time.

Bacterial pan- and core-genomes – a comparative genomics approach based on bidirectional blast

Wollherr, A.*, Liesegang, H.

Göttingen Genomics Laboratory, Insitute for Microbiology and Genetics, Grisebachstraße 8, 37077 Göttingen

*Corresponding author Email: <awollhe@gwdg.de>

Conserved gene clusters within a species build the core genome encoding the basic functions of a species whereas highly variable parts, called the pan genome, determine the abilities decisive for niche adaption. Thus the identification of gene clusters and their assignment to the core resp. the pan-genome is essential for the understanding of the lifestyle of organisms.

With our approach the core and pan genome regions of related genomes can be identified. Therefore a softwaretool called BiBaG is implemented. BiBaG represents a pipeline which consists of a bidirectional blast receiving orthologous genes followed by the needleman-wunsch alignment used to evaluate the putative orthologs based on global identity scores.

The method was applied in comparative genomics on *Bacilli,* methanogenic archaea and sulfat reducing bacteria. The produced data revealed interesting insights in the genome organization and the ecological adaption of the *B. subtilis* group compared to the *B. cereus* group.

Natural Similarity Measures between Position Frequency Matrices with an Application to Clustering

Utz J. Pape, Sven Rahmann*, and Martin Vingron

Bioinformatics for High-Throughput Technologies
Computer Science 11, TU Dortmund, D-44221 Dortmund, Germany

*Corresponding author Email: Sven.Rahmann@tu-dortmund.de

Transcription factors (TFs) play a key role in gene regulation by binding to target sequences. In silico prediction of potential binding of a TF to a binding site is a well-studied problem in computational biology. The binding sites for one TF are represented by a position frequency matrix (PFM). The discovery of new PFMs requires the comparison to known PFMs to avoid redundancies. In general, two PFMs are similar if they occur at overlapping positions under a null model. Still, most existing methods compute similarity according to probabilistic distances of the PFMs. Here we propose a natural similarity measure based on the asymptotic covariance between the number of PFM hits incorporating both strands. Furthermore, we introduce a second measure based on the same idea to cluster a set of the Jaspar PFMs.

We show that the asymptotic covariance can be efficiently computed by a two dimensional convolution of the score distributions. The asymptotic covariance approach shows strong correlation with simulated data. It outperforms three alternative methods. The Jaspar clustering yields distinct groups of TFs of the same class. Furthermore, a representative PFM is given for each class. In contrast to most other clustering methods, PFMs with low similarity automatically remain singletons.

A website to compute PFM similarity and clusters, including source code and supplementary material, is available at `http://mosta.molgen.mpg.de/`. Details can be found in [1].

References

1. U. J. Pape, S. Rahmann, and M. Vingron. Natural similarity measures between position frequency matrices with an application to clustering. *Bioinformatics*, 24(3):350–357, Feb 2008.

A discriminative approach for de-novo motif discovery

Jens Keilwagen°*, Jan Grau°, Ivo Grosse, Marc Strickert and Stefan Posch

IPK Gatersleben
Institute of Computer Science, Martin Luther University Halle–Wittenberg

*Corresponding author Email: Jens.Keilwagen@ipk-gatersleben.de
°These authors contributed equally

DNA-binding molecules play an important role in the regulation of cellular processes. These molecules, e.g. transcription factors (TFs), bind to short stretches of DNA and influence the rate of transcription. Hence, the de-novo discovery of these DNA-motifs is one of the keys to understand gene regulation.

In many cases, TFs are functional only when binding coordinately with other TFs, forming cis-regulatory modules (CRMs). Positional preference is another key feature to distinguish functional from non-functional binding sites for many TFs. Finally, the search for short over-represented motifs often leads to false positives. However, not the most frequently occurring motifs are the most interesting ones, but those explaining best differential regulation, e.g. between groups of genes. Existing algorithms address subsets but not all of these issues.

Here, we propose a new algorithm for de-novo motif discovery that optimizes a *discriminative* objective function in order to discover differential motifs or CRMs. It detects existing positional preferences as well as dependencies between binding sites of different TFs. The proposed algorithm does not require the user to exactly specify the length of a motif, but allows for an internal adaptation.

The proposed algorithm improves the de-novo discovery of DNA-motifs on a realistic artificial dataset, especially when positional preference is strong. On this dataset, the algorithm discovers motifs that could not be detected by the MEME algorithm [1], while showing a higher accuracy than DEME [2], another discriminative algorithm.

References

1. Timothy L. Bailey and Charles Elkan. Fitting a mixture model by expectation maximization to discover motifs in biopolymers. In *Proceedings of the Second International conference on intelligent systems for molecular biology*, 1994.

2. Emma Redhead and Timothy Bailey. Discriminative motif discovery in DNA and protein sequences using the DEME algorithm. *BMC Bioinformatics*, 8(1):385, 2007.

Probabilistic Arithmetic Automata and their Applications

Tobias Marschall, Inke Herms, and Sven Rahmann*

Bioinformatics for High-Throughput Technologies
Computer Science 11, TU Dortmund, D-44221 Dortmund, Germany

*Corresponding author Email: Sven.Rahmann@tu-dortmund.de

A common problem in biological sequence analysis is the discovery of overrepresented patterns (motifs) in sequences with the typical sub-problem "What is the probability of observing at least n matches of a given motif candidate by chance?".

We have introduced the concept of *probabilistic arithmetic automata (PAAs)* at CPM'08 [3] and show here how it paves the way for a dynamic programming approach to exact pattern matching statistics. It can also be applied to compute fragment statistics of protein cleavage reactions [2] and *seed sensitivity* for pairwise alignment [1], to be presented at WABI'08.

Definition 1 (PAA). A *probabilistic arithmetic automaton* is an 8-tuple $(Q, T, q_0, N, n_0, E, \theta = (\theta_q)_{q\in Q}, \pi = (\pi_q)_{q\in Q})$, where Q is a finite set of states, $T : Q \times Q \to [0, 1]$ is a transition function, i.e., $(T(p, q))_{p,q\in Q}$ is a stochastic matrix, $q_0 \in Q$ is called *start state*, N is a finite set called *value set*, $n_0 \in N$ is called *start value*, E is a finite set called *emission set*, each $\theta_q : N \times E \to N$ is an operation associated with the state q, and each $\pi_q : E \to [0, 1]$ is an emission distribution associated with the state q.

At first, the automaton is in its start state q_0, as for a classical deterministic finite automaton (DFA). In a DFA, the transitions are triggered by input symbols. In a PAA, however, the transitions are purely probabilistic; $T(p, q)$ gives the chance of going from state p to state q. While going from state to state, a PAA performs a chain of calculations on a set of values N. In the beginning, it starts with the value n_0. Whenever a state transition is made the entered state, say state q, generates an emission according to the distribution π_q. The current value and this emission are then subject to the operation θ_q, resulting in the next value. Let $\{Y_k\}_{k\in\mathbb{N}_0}$ denote the automaton's state process and $\{V_k\}_{k\in\mathbb{N}_0}$ the *value process*, that means the sequence of values resulting from the chain of performed calculations.

PAAs thus provide a formalization of computations on (conditional) probability distributions. In particular joint state-value probabilities $\mathbb{P}(Y_k = y, V_k = v)$ can be computed exactly, for a variety of binary operations on values and emissions (frequently, addition or maximization of natural numbers, such as match counts).

The poster presents applications to Prosite patterns, and both threshold-based and fuzzy PWM representations of transcription factor binding sites.

References

1. I. Herms and S. Rahmann. Computing alignment seed sensitivity with probabilistic arithmetic automata. In K. Crandall and J. Lagergren, editors, *Algorithms in Bioinformatics (WABI'08)*, LNCS. Springer, 2008. In press.
2. H.-M. Kaltenbach, S. Böcker, and S. Rahmann. Markov additive chains and applications to fragment statistics for peptide mass fingerprinting. In T. Ideker and V. Bafna, editors, *Systems Biology and Computational Proteomics*, volume 4532 of *Lecture Notes in Computer Science*, pages 29–41. Springer, 2006.
3. T. Marschall and S. Rahmann. Probabilistic arithmetic automata and their application to pattern matching statistics. In P. Ferragina and G. Landau, editors, *Combinatorial Pattern Matching (CPM'08)*, volume 5029 of *LNCS*, pages 95–106. Springer, 2008.

SQUIRREL: Domain decomposition based on vectorial structure representation

Florian Teichert, Ugo Bastolla and Markus Porto[*]

TU Darmstadt, Institut für Festkörperphysik, Darmstadt, Germany

*Corresponding author Email: porto@fkp.tu-darmstadt.de

Decomposing protein structures into their constituent domains is still a challenging task for computational biology, both because the definition of domain is not clear-cut and because of the inherent complexity of protein structures. We present here SQUIRREL, a new method for automatic domain decomposition based on a vectorial representation of protein structures. By applying our method to a benchmark set of protein structures that have been previously identified as `difficult' for established methods, we show that our new approach yields decompositions that are in very good agreement with human experts' assessments. To get a more quantitative measure for the quality of domain decompositions by our method, we furthermore perform a large scale domain decomposition of a representative subset of the Protein Database using our method as well as other established ones and quantify the quality of domain decomposition by recurrence.

JSTACS – A Java framework for statistical analysis and classification of biological sequences

André Gohr°, Jens Keilwagen°, Jan Grau°*, Michael Seifert, Michaela Mohr, Stefan Posch and Ivo Grosse

Institute of Computer Science
Martin Luther University Halle–Wittenberg

*Corresponding author Email: grau@informatik.uni-halle.de

°These authors contributed equally

Sequence analysis is one of the major subjects of bioinformatics. Several existing libraries combine the representation of biological sequences with exact and approximate pattern matching as well as alignment algorithms. We present Jstacs, an open source Java library which focuses on the statistical analysis of biological sequences instead. Jstacs comprises an efficient representation of sequence data and provides implementations of many statistical models, such as position weight matrix (PWM) models, weight array models (WAMs), Markov models of arbitrary order, or Bayesian networks. For learning parameters of statistical models Jstacs provides generative and discriminative approaches, including maximum likelihood, maximum a-posteriori, maximum conditional likelihood, or maximum supervised posterior. Jstacs allows to combine models to classifiers e.g. for distinguishing donor splice sites from non-splice sites. Using Jstacs, classifiers can be assessed and compared on test datasets or by cross-validation (CV) experiments evaluating several performance measures, e.g. sensitivity, false positive rate, positive predicitive value, or area under ROC curve.

The strictly object-oriented design of Jstacs allows for a straightforward extension of abstract base classes to implement additional statistical models or classifiers, sophisticated assessment schemes, specialized optimization techniques, or upcoming approaches for parameter learning. The different levels of abstraction and many default implementations greatly reduce a programmer's effort. For instance, defining a new generative statistical model basically requires to implement three obligatory and some optional methods, e.g. for saving it as an XML-representation using the XML-parser of Jstacs. As a further example, the comparison of generatively as well as discriminatively trained PWMs and WAMs utilizing a 10-fold CV on a foreground and a background dataset takes less than 20 lines of code from loading the input data to printing the performance measures for the four studied classifiers.

Hence, Jstacs has the capability to support the development and comparability of new statistical models in bioinformatics.

Molecular Fragmentation Query Language for Shotgun Lipidomics

Ronny Herzog*, Dominik Schwudke and Andrej Shevchenko
MaxPlanckInstitute-CBG; Pfotenhauerstr. 108; 01307 Dresden
*Email: herzog@mpi-cbg.de

Mass spectrometry is the method of choice to analyze the molecular composition of cellular lipidomes. Resulting large datasets of tandem mass spectra are interpreted in common by comparison of experimental spectra versus reference spectra of component databases. This approach is inefficient in shotgun lipidomics, in which most of MS/MS spectra are taken from mixtures of several isobaric lipid components. Furthermore, database-bound data interpretation is limited by the number of its reference entries. The alternative de novo interpretation is laborious, time consuming and requires high expertise. We have developed the Molecular Fragmentation Query Language (MFQL) to automate de novo interpretation of large lipidomic datasets.

Results

Instruments raw data was converted to *.dta or *.mzXML files by the vendor's software or tools of the Trans-Proteomic Pipeline (http://tools.proteomecenter.org/software.php), respectively.
Mass spectra from shotgun lipidomic experiments of single biological samples, are imported into a comprehensive, tree-like structured mass spectra database. Single scan spectra are averaged and aligned according to several instrument-specific settings, such as mass resolution and mass accuracy.
MFQL is designed similar to SQL. So, Molecular fragmentation patterns are formalized as queries to identify and quantify the lipid species in the experimental spectra database. In particular, an unlimited number of precursor and neutral loss scans (Schwudke 2006), and boolean combinations thereof, can be emulated.
Currently, fragmentation pathways of major lipid classes have been formalized that allow a fast "click-and-find-lipids" interpretation of shotgun datasets.
To our knowledge, LipidX is the first software for the quantitative interpretation of large shotgun lipidomic (metabolomic) datasets with a query language. We will present several applications of LipidX for shotgun profiling of complex eukaryotic lipidomes.

GabiPD: The Gabi Primary Database - a plant integrative 'omics' database

Diego Mauricio Riaño-Pachón, Axel Nagel, Robert Wagner, Jost Neigenfind, Elke Weber, Svenja Diehl, Birgit Kersten*

GabiPD team, Bioinformatics group, Max Planck Institute of Molecular Plant Physiology

*Corresponding author Email: gabipd@mpimp-golm.mpg.de

The GABI Primary Database, GabiPD, was established eight years ago in the frame of the German initiative for Genome Analysis of the Plant Biological System (Genomanalyse im biologischen System Pflanze, GABI). The main goal of GabiPD is to collect, integrate, visualize and link primary information from GABI projects. GabiPD, in contrast to other plant databases constitutes a repository and analysis platform for a wide array of heterogeneous data arising from high-throughput experiments in several plant species. Currently, data from different 'omics' fronts are incorporated in GabiPD (i.e., genomics, transcriptomics, metabolomics, proteomics), originating from 14 different model or crop species. We have developed the concept of GreenCards for text based retrieval of all data types in GabiPD (e.g., clones, genes, mutant plant lines, markers). All data types are pointing to the central Gene's GreenCard, where gene information is integrated from genome annotation projects. Within the Gene's GreenCards links to all GabiPD data related to the corresponding genes as well as cross references to large UniGene sets from NCBI and to useful gene-based external data bases are displayed. A collection of 400000 ESTs from different species, generated in different GABI projects, is made publicly available though GabiPD. These ESTs have been cross referenced to UniGene sets from NCBI and to sequences from different plant genome projects, in an effort to ease the transfer of functional information. The centralized Gene's GreenCard also allows visualizing ESTs aligned to annotated transcripts as well as identified protein domains and gene structure. Moreover GabiPD makes available interactive genetic maps from *Solanum tuberosum* (potato) and *Hordeum vulgare* (barley). Gene expression data in GabiPD can be visualized through MapManWeb, the web interface of MapMan. Access to the data in GabiPD is provided via either the web interface (http://www.gabipd.org/) or webservices that are currently available for Arabidopsis-related information.

GenColors: Annotation and Comparative Genomics Made Easy

Marius Felder, Alessandro Romualdi, Gernot Glöckner, Matthias Platzer and Jürgen Sühnel*

Biocomputing Group, Leibniz Institute for Age Research - Fritz Lipmann Institute (FLI), Beutenbergstr. 11, D-07745 Jena, Germany

*Corresponding author Email: jsuehnel@fli-leibniz.de

GenColors is a web tool initially aimed at the annotation and analysis of prokaryotic genomes with an emphasis on genome comparison. As a new feature the system has now been adapted to handle also eukaryotic genomes.

Results

GenColors is a web-based software/database system initially aimed at an improved and accelerated annotation of prokaryotic genomes making extensive use of genome comparison (Romualdi *et al.*, Bioinformatics 2005; Romualdi *et al.*, Methods Mol. Biol. 2007). It offers a seamless integration of data from ongoing sequencing projects and annotated genomic sequences obtained from GenBank. With GenColors dedicated genome browsers containing a group of related genomes can be easily set up and maintained. The tool has been efficiently used for *Borrelia garinii* (Glöckner *et al.*, Nucleic Acids Res. 2004; Glöckner *et al.*, BMC Genomics 2006) and is currently applied to a number of other ongoing genome projects on *Legionella*, *Pseudomonas* and *E. coli* genomes. Examples for freely accessible GenColors-based dedicated genome browsers are the Spirochetes Genome Browser SGB (sgb.fli-leibniz.de), the Photogenome Browser CGB (cgb.fli-leibniz.de) and the Enterobacter Genome Browser ENGENE (engene.fli-leibniz.de).

The system has now been adapted to handle also eukaryotic genomes. A first application of this feature is the ongoing annotation and analysis of two fungal species.

Another GenColors-based tool is the Jena Prokaryotic Genome Viewer - JPGV (jpgv.fli-leibniz.de). Contrary to the dedicated browsers it offers information on almost all finished bacterial genomes. Currently, it includes 1140 genomic elements of 293 species.

Information Retrieval in Life Science: A Novel Search Engine

M. Lange, J. Bargsten, M. Klapperstück, P. Schweizer , U. Scholz and K. Spies*

Institut of plant genetics and crop plant research (IPK)
Corrensstrasse 3
06466 Gatersleben

*Corresponding author Email: spies@ipk-gatersleben.de

Motivation

Seeking information across life science databases is a time consuming task in dry lab research. Searching scientific databases effectively necessitates the use of contemporary software to locate desired and meaningful information. Web search engine like Google provide a fast search in common web content and provide popularity ranking. Retrieval systems for scientific data, like ENTREZ, provide powerful query interfaces on top of integrated life science databases. Neither keyword search, popularity ranking of web pages nor query engines and retrieval systems are a reliable information research system.

Method

This gap is filled by a new kind of life science search engine that combines methods from web search engines, data retrieval systems and text mining. In order to extract relevant and qualitative knowledge, our approach combines a fast search engine and a novel context based relevance ranking with methods from artificial intelligence. With the support of user profiles and automatic synonym expansion, we are able to deliver the most relevant database entries matching best the query context. Each found database entry is described by a 9 dimensional feature vector which is translated into a scalar relevance score.

Results

Comparisons between our ranking and manual ranked use-cases show very promising values of precision (100%) and recall (over 85%). In order to increase the ranking quality we support tracking of user interactions and user specific result evaluation like voting for example. The system learns automatically from user provided information and integrates this into following rankings of queries. This new method of data seeking combines a clean, powerful and easy to use human computer interface with a rule based, context sensitive ranking system. The result is a search engine, which brings a new quality into the scientific dry lab research. System is available by request to the authors.

Computer Aided Quantitative Evaluation of Non-alcoholic Fatty Liver Disease

Štěpán Holinka* and Daniel Smutek

3rd Medical Department, 1st Faculty of Medicine, Charles University in Prague

*Corresponding author Email: stepanholinka@gmail.com

Non-alcoholic fatty liver disease or non-alcoholic steatohepatitis (NASH) is a progressive liver disease, which poses risk of advanced liver failure. The disease is related to fat deposition in liver tissue. Quantitative evaluation of amount of fat in the liver tissue is currently done mainly invasively by liver biopsy. This examination is the gold standard for diagnosis, though it also raises the risk of mortality. There are also non-invasive medical imaging techniques such as ultrasonography, computerized tomography and magnetic resonance which are able to demonstrate fat in the liver tissue. The ultrasonography is relatively cheap and quick diagnostic technique, but it gives mainly qualitative assessment based on a physician's experience. In this paper we present a methodology for quantitative evaluation of the amount of fat in the liver tissue by automatic texture analysis of B-mode sonographic images. Sonographic image data of patients with NASH are acquired and evaluated by three hepatology physicians into four groups, similarly as during routine investigation. The evaluation is based on the physicians' experience and it is in accordance with the fat amount in patients' liver. As a confirmation method the liver biopsy is used. This objective examination allows dividing the subjects also into four groups by using osmium tetroxide method. The most important part of this method is a feature selection. We use the feature set comprising Haralick's co-occurrence features and Muzzolini's spatial features. This set of texture features proved to be very successful in evaluating diffuse changes in ultrasound imaging in other parenchymal organs [1].

References

1. Smutek D., Šára R., Sucharda P., Tjahjadi T., Švec M. : Image texture analysis of sonograms in chronic inflammations of thyroid gland. *Ultrasound in Medicine and Biology*, 29(11):1531–1543, 2003.

Novel image analysis applications for liver fibrosis measurements

Kaczmarek E.[1*], Jagielska J.[1], Chmielewski M.[2]

[1]Lab. of Morphometry and Medical Image Analysis and [2]Dept. of Gastroenterology at Poznan University of Medical Sciences.

*corresponding author Email: elka@ump.edu.pl

The aim of our contribution was to present a 3D image analysis method for segmentation and quantification of specific structures in liver fibrosis cases. Progression of liver fibrosis has usually been evaluated by liver biopsy using insensitive descriptive semi quantitative score systems. An alternative to this is to perform quantitative measurements. However, its accuracy remains sometimes controversial because of sampling variability caused by the small size of biopsy samples and the heterogeneity of liver fibrosis. Among noninvasive imaging methods, elastography has been shown to be a reliable method. It is based on the observation that fibrosis leads to increased tissue stiffness. The purpose of our study was also to compare results of ultrasound elastography with histomorphometrical evaluation of microphotographs.

Material and methods

Needle biopsies obtained in patients with liver fibrosis were routinely processed. Sections of each biopsy stained with Sirius red were used for measurements. Microphotographs were acquired by a digital microscope running under Motic Images software. Our method based on spatial visualization of specific colors was performed to extract fibrous areas. Pixels representing the specific colors were converted into voxels by introducing their brightness as third dimension and revealed on a spatial view by reducing the scenery behind to a background.

Ultrasound elastography images were registered with a EUB-6500HV ultrasound system with a linear transducer EUP-L73S. Elastography measurements were performed on different liver regions, mostly on the right lobe of the liver with a constant pressure.

Moreover, the activation of immune system plays an important role in the development of liver injury, therefore a 3D segmentation of cytokines from microphotographs will be also presented.

Results

The average degree of fibrosis based on biopsy examination ranged from 6.4% to 37.1% while in elastography from 5.4% to 37.4%. The elastographic results were significantly correlated with measurements performed in microphotographs ($r=0.754$, $p<0.05$).

A new Image Analysis Approach in Systems Biology for Investigating Infection Processes of human pathogen Fungi

F. Mech[1]*, S. Krause[2], A. Brakhage[1]

[1]Leibniz Institute for Natural Product Research and Infection Biology - Hans-Knöll-Institute (HKI), D-07745 Jena, Germany

[2]Friedrich Schiller University of Jena, D-07743 Jena, Germany

*Corresponding author Email: franziska.mech@hki-jena.de

A number of studies on human-pathogen fungi, especially *Candida albicans* and *Aspergillus fumigatus*, demonstrate the variety of virulence strategies of those parasites. A new approach to understand the interactions between host and pathogen shall be established. Time laps fluorescence microscopy and PET-CT (positron-emission topography combined with computer tomography) are used to generate image and movie files visualizing host-pathogen interaction in laboratory animals or in vitro experiments [1]. Due to the huge amount of data the analysis of these movies requires an almost automatically image analysis. Definite patterns (objects, like macrophages, neutrophiles and conidia) and events (adhesion, colonization, infection, phagocytosis) have to be recognize efficient and robust. There are special tools to cope with the challenge. For this project software from Definiens is used. Rulesets have to be developed and applyed to the movies for quantifying the pre- or absence of the mentioned criterias as well as their characteristics automatically. After that the obtained data is used to validate game-theoretically approaches for mathematical modelling of the infection events.

[1] J. Behnsen, P. Narang, M Hasenberg, F. Gunzer, U. Bilitewski, N. Klippel, M. Rohde, M. Brock, AA. Brakhage, M. Gunzer: Environmental Dimensionality Controls the Interaction of Phagocytes with the Pathogenic Fungi *Aspergillus fumigatus* and *Candida albicans* (2007) *PloS Pathogens* Vol 3, Issue 2, 138-151

A New Approach in Systems Biology of Infection Processes of Human Pathogen Fungi

S. Hummert[1]*, F. Mech[1], A. Schröter[2], S. Schuster[2], R. Guthke[1]

[1] Department Molecular and Applied Microbiology, Leibniz Institute for Natural Product Research and Infection Biology, Hans-Knöll-Institute Beutenbergstr. 11a, D-07745 Jena, Germany

[2] Bioinformatics, Institute for Biology and Pharmacy, Friedrich Schiller University (FSU) Jena, Ernst Abbe Platz 2, D-07743 Jena, Germany

*Corresponding author Email: sabinehu@minet.uni-jena.de

Motivation

A number of studies on human-pathogen fungi, especially *Candida albicans* and *Aspergillus fumigatus*, demonstrate the variety of virulence strategies of those parasites. A new approach to understand the interactions between host and pathogen shall be established. Time laps fluorescence microscopy and PET-CT (positron-emission topography combined with computer tomography) are used to generate image and movie files visualizing host-pathogen interaction in test animals or in vitro experiments [B07]. IT processes shall be developed for an automatic evaluation of the huge amount of those data. Certain patterns (objects like macrophages or conidia) and events (e.g. adhesion, colonization, infection or phagocytosis) have to be detected reliably.

Methods

Mathematical methods shall be used to model the infection events spatio-temporally. A first game-theoretical approach of a 'conflict' versus 'compromise' game [R91] is studied and extended to show the transition from commensalism (in the yeast form of *Candida*) to virulence (in the hyphal form). This model shows that being aggressive is always a Nash equilibrium, while being cooperative is a Nash equilibrium only if virulence costs for the parasite and resistance costs for the host are too high. At equal costs, the compromise-system gives a higher fitness for both.

On the other hand Conway's Game of Life is extended to more then one organism. So virtual macrophages and conidia can be simulated and artificial videos created. In future the game rules shall be created by machine learnig techniques.

References

B07. Behnsen J, Narang P, Hasenberg M, Gunzer F, Bilitewski U, Klippel N, Rohde M, Brock M, Brakhage AA, Gunzer M: Environmental Dimensionality Controls the Interaction of Phagocytes with the Pathogenic Fungi Aspergillus fumigatus and Candida albicans. In: PloS Pathogens 3(2) 2007, 138–151.

R91. Renaud F, de Meeûs F: A simple model of host-parasite evolutionary relationships. Parasitism: compromise or conflict? In: J. Theor. Biol. 152, 1991, 319–327.