PREDICTION AND ANNOTATION OF MOLECULAR FUNCTION

Chairs: Nir Ben-Tal and Frederick Roth

D-1. Validating the structural information exchange within the domain Fyn SH2

Huculeci R (1,2,), Buts L (1,2), Rousseau F (3,4), Schymkowitz J (3,4), van Nuland N (1,2), Lenaerts T (5,6)*

In recent years, a variety of in silico methods have been developed that try to predict, starting from sequential or structural information, the long-range (allosteric) effects of peptide binding on protein structures. These predictions are highly relevant since they expose the functionally relevant parts of the protein, allowing one to gain scientific insight into highly important regions and their possible relations to known diseases.

Materials and Methods

Our own approach uses Shannon's information theory to quantify the conformation coupling between the residue sidechains [1]. When comparing the mutual information profiles of the bound and unbound states of a domain, regions can be identified that may play a role in signal propagation. In order to validate our predictions we have initiated an extensive NMR study to analyze the structural and dynamic features of both isolated and peptide bound Fyn SH2.

Results

In a first stage two new structural ensembles, representing the two unbound and bound state, have been determined and these were used to predict the network of residues affected by binding. In a second stage, the binding dynamics are experimentally examined through an NMR analysis of the methyl dynamics. This analysis provides the experimental validation for our predictions. In this poster we will present the results that are obtained so far : novel Fyn SH2 structures and the newly predicted and experimental results

Discussion

This work clearly shows that the predicted residues are indeed involved in the information processing by the domain structure. Moreover, it provided the first experimental proof of our results. As such it opens the road to an new understanding of domain structure and domain engineering in the longer term.

URL

http://www.ulb.ac.be/di/map/tlenaert/

Presenting Author

Tom Lenaerts (tlenaert@ulb.ac.be)
MLG, DI, Université Libre de Bruxelles

Author Affiliations

1 Structural Biology Brussels, Vrije Universiteit Brussel, Pleinlaan 2, 1050 Brussel, Belgium 2 Department of Molecular and Cellular Interactions, VIB, Pleinlaan 2, 1050 Brussel, Belgium 3 Switch, VIB,Pleinlaan 2, 1050 Brussel, Belgium 4 Vrije Universiteit Brussel, Pleinlaan 2, 1050 Brussel, Belgium 5 MLG, Département d'informatique, Université Libre de Bruxelles, Boulevard du Triomphe CP212, 1050 Brussel, Belgium 6 Department of Computer Science, Vrije Universiteit Brussel, Pleinlaan 2, 1050 Brussel, Belgium

Acknowledgements

The authors acknowledge the support of the FWO, FNRS and the VIB for this project

D-2. Discriminant functions for classification and prediction of T1SS, T3SS, T4SS and T6SS secreted proteins from proteobacteria

Kampenusa I (1,2,), Zikmanis P (2)*

There are at least four systems being responsible for leaderless secretion in proteobacteria. Nevertheless, exoproteins experimentally annotated to the definite secretion type, still remain as a trifling part of sequenced exoproteome. The available prediction tools can only attribute leaderless proteins as a whole group without specifications in respect to appropriate secretion types. Recently some computational approaches have been proposed to identify substrates of leaderless secretion within the bacterial exoproteome, however representing the very narrow range of organisms.

Materials and Methods

Data set: 111 proteins from 30 genera of proteobacteria annotated as secreted by T1SS, T3SS, T4SS or T6SS (Swiss-Prot/TrEMBL ID numbers are summarized in: Kampenusa and Zikmanis 2010). The frequencies of amino acids and codon usage were computed by ExPASy/ProtParam and Emboss/Cusp tools, respectively. Multiple Discriminant analysis and Press's Q test were performed by SPPS 11.0 for Windows. The prediction success of obtained sets of predictor variables (i.e., the classifiers) were compared by means of McNemar's test.

Results

Frequencies of 24 codons, partially representing 16 amino acids (C D G I K L M N P Q R S T V W Y), and 10 amino acids (G K M N P Q R W Y E) were selected as two sets of predictor variables from the sets of codon usage and amino acid frequencies for multiple-4-group discrimination of T1SS, T3SS, T4SS and T6SS secreted proteins (Q 309.43 & 272.08, $p< 0.001$; respectively). The both classifiers exhibit comparable error rates ($p>0.05$, McNemar's test) as well as the equal prediction success (81%) for the affiliation of proteins from independent testing-sets.

Discussion

The 10 sequences from the training set of proteins being misclassified by the both classifiers exhibit an opposite trend of compositional characteristics as compared to the whole set of proteins. Therefore it is likely, that restrictions of proposed discriminant functions in respect of an adequate classification reflect somehow different evolutionary routes of the protein sequences. Nevertheless, the both classifiers can be used as either independent or complementary tools to specify the substrates of leaderless secretion with potent more widespread applications outside the proteobacteria.

Presenting Author

Ināra Kampenusa (inara.kampenusa@gmail.com)
University of Latvia, Faculty of Biology, Institute of Microbiology and Biotechnology

Author Affiliations

1 Faculty of Biology, University of Latvia 2 Institute of Microbiology and Biotechnology, University of Latvia

Acknowledgements

Support for the research has been granted by ESF project “Support of doctoral studies at the University of Latvia”.

D-3. Computing fragmentation trees from tandem mass spectrometry data

Rasche F (1,), Svatos A (2), Böcker S (1)*

Plants and other organisms produce enormous numbers of unknown metabolites. Identification of such substances is a common task metabolomics and pharmaceutical chemistry. Tandem mass spectrometry is used to achieve this, but computational analysis of the spectra is still in its infancy. Here we introduce fragmentation trees representing the fragmentation reactions of the molecule. We also present a method for the calculation of hypothetical fragmentation trees. This method does not require a molecular structure or mass spectral database.

Materials and Methods

To calculate the tree we first generate a fragmentation graph containing all fragmentation trees that are in accordance with the measured spectra. In the second step, this graph is scored using properties such as peak intensities, mass deviations, typical neutral losses, etc. Finally, we search the highest-scoring fragmentation tree inside the graph. Since we want to explain every peak only once, we have to apply restrictions that render the problem NP-hard. A fixed parameter tractable algorithm is used to solve the problem in reasonable running time.

Results

We tested the method on spectra of 180 compounds from 3 different instruments. Mass spectrometry experts evaluated the resulting trees. They considered 79% of the fragmentation reactions to be "correct". The experts used multiple MS spectra of some compounds to support their assignment. Additionally, we used the commercial software MassFrontier (MF) to assess our peak annotation. MF annotates a spectrum when given a compound structure. MF and our tool assigned the same molecular formula to 97% of the peaks annotated by MF, but our tool annotates four times as many peaks as MF.

Discussion

We have demonstrated that the calculated fragmentation trees agree well with expert-knowledge and rule-besed fragmentation prediction. Our method is fully automated and does not require any previous knowledge about the compound. Fragmentation trees reveal more about the molecular structure than the spectrum alone. Therefore, in future, we might discover structural similarities between known and unknown compounds by comparing their fragmentation trees. We are currently developing a method for this comparison.

URL

http://bio.informatik.uni-jena.de/software/starburst

Presenting Author

Florian Rasche (florian.rasche@uni-jena.de)
Friedrich-Schiller-University Jena

Author Affiliations

(1) Chair of Bioinformatics, Friedrich-Schiller-University Jena, Germany (2) Research Group Mass Spectrometry, Max Planck Institute for Chemical Ecology, Jena, Germany

Acknowledgements

The Max Planck Society supported this work financially.

D-4. Computational identification of new miRNAs in genomic regions associated with psychiatric diseases

Aelterman B (1,2,), De Rijk P (1,2), Del-Favero J (1,2)*

miRNAs are small (~22 nt) noncoding RNAs that are important regulators of gene expression. The latest release of miRBase (15.0) contains 940 human miRNA sequences of which several are already implicated in human diseases such as Tourette's symdrome, autism, schizophrenia, ...

Materials and Methods

We developed a miRNA prediction pipeline that integrates different programs to predict novel miRNAs in genomic regions associated with psychiatric diseases. In total we used 8 different miRNA prediction programs for the construction of our miRNA prediction pipeline. Predictions of all programs were clustered to form consensus predictions and a score was allocated to each prediction based on the number of different programs supporting the prediction. We used available Next Generation Sequencing data to validate the obtained predictions.

Results

We ran our pipeline on 15 genomic regions, all of which are associated or linked to psychiatric diseases. The total size of these regions is approximately 1,5% of the human genome. Over 300.000 miRNAs were predicted with prediction scores ranging from 1 to 7. To add experimental evidence to the predictions, we developed an in silico validation platform based on sequence similarity between the miRNA predictions and sequence reads from RNA-Seq data available via NCBI's Sequence Read Archive. When selecting all miRNA predictions with a prediction score above 5 (n=163) we were able to show that 21 predicted miRNAs were present in the RNA-seq data.

Discussion

We were able to predict novel miRNAs and could validate 21 novel miRNAs. As these are located in genomic regions associated with schizophrenia or bipolar disorder, they are potentially genetic risk factors for these diseases. Since we validated 21 novel miRNAs from 1,5% of the human genome it is likely that many more miRNAs are to be discovered.

Presenting Author

Bart Aelterman (bart.aelterman@molgen.vib-ua.be)
Applied Molecular Genomics Group, VIB Department of Molecular Genetics, University of Antwerp

Author Affiliations

(1) Applied Molecular Genomics group, VIB Department of Molecular Genetics, VIB, Antwerp, Belgium (2) University of Antwerp (UA), Antwerp, Belgium

Acknowledgements

The authors would like to thank IWT, VIB and UA for their support.

D-5. CatANalyst: a web server for predicting catalytic residues

Cilia E (1,2,), Passerini A (1)*

The identification of the residues involved in the catalytic processes is a key step for fully characterizing an enzyme function. Understanding the molecular mechanisms of protein functioning is difficult even when its 3D structure is known. The process is time consuming and goes through several hypothesis formulation and verification steps by targeted experiments (e.g. site-directed mutagenesis). Machine learning approaches can significantly speed up the whole process by suggesting candidates to be experimentally verified and potentially allowing automatic annotation of functional residues.

Materials and Methods

CatANalyst takes a FASTA sequence or a PDB structure and builds a residue representation that is fed to an SVM classifier. Sequential features are based on multiple alignment profiles. Structural features are extracted from spherical regions around residues and include features like solvent accessibility, relative position on the protein surface, hydrogen bonds, secondary structure. CatANalyst outputs the probability of being catalytic for residues exceeding an adjustable threshold. All residues are also rendered with different colors and sizes, reflecting how likely they are to be catalytic.

Results

The structured-based CatANalyst was previously shown to achieve consistent improvements over existing approaches on several benchmark datasets. The online version of the predictors was trained on a dataset of 137,116 residues of 505 enzymes from PDBselect and deposited in PDB before 2008. Results on a test set of 88 newly deposited enzymes are consistent with the cross-validation results obtained on the benchmark datasets. As a case study we also tested CatANalyst on a predicted model of an amidase: predictions include the known catalytic residues and also putative ones reported in literature.

Discussion

CatANalyst provides both sequence- and structure- based functional residue predictions. The classifier shows a slight positive bias on residues having a high catalytic propensity that bind catalytic cofactors. These residues do not have a direct role in the catalysis according to the CSA annotation. On the other hand, information on nearby heterogens helps in identifying functional residues with low catalytic propensities. Planned extensions include the joint prediction of both catalytic and binding site residues and the collective prediction of residues belonging to the same catalytic site.

URL

http://catanalyst.disi.unitn.it/

Presenting Author

Elisa Cilia (elisa.cilia@iasma.it)
Fondazione Edmund Mach - Istituto Agrario S. Michele all'Adige

Author Affiliations

(1) Department of Engineering and Computer Science, University of Trento via Sommarive, 14, I-38123 Trento, Italy {cilia,passerini}@disi.unitn.it (2) Fondazione Edmund Mach - Istituto Agrario S. Michele all'Adige via Edmund Mach 1, I-38010 S. Michele all'Adige (TN), Italy elisa.cilia@iasma.it

Acknowledgements

Provincia Autonoma di Trento

D-6. Functional characterization of Parkinson by high-throughput data analysis with l1l2 regularization

Squillario M (1,2,), Masecchia S (2), Barla A (2)*

Parkinson disease (PD) is a neurodegenerative disorder that impairs the motor skills at the onset and the cognitive and the speech functions successively. Like other neurological diseases, once the first symptoms appear, a great loss of neurons has already occurred and so far no clinical tests are able to early diagnose the disease long time before the first symptoms appear. To develop a reliable clinical test, it is necessary to understand the disease at a molecular level. In this context, we analyze microarray data by means of the l1l2 regularization framework.

Materials and Methods

From the statistical viewpoint, we use the l1l2 regularization framework, a very interesting method when dealing with high-throughput data, given its ability of providing sparse predictive models. Based on soft thresholding, this approach is consistent, multivariate and capable of dealing with correlated variables. The double loop of cross validation guarantees an unbiased result. We also characterize the list of relevant probesets with a functional analysis performed by WebGestalt, a toolkit that allows for enrichment analysis on the Gene Ontology and other digital repositories.

Results

The feature selection analysis results in a 64% prediction accuracy, associated to a signature composed by 378 selected probesets for the maximum correlation allowed. The functional analysis shows that many of the genes are involved in pathways related to the immune system, to some particular cell basic processes (apoptosis, cell cycle, motility, communication) but also to the nervous system, to the metabolism and to some diseases/infections. Moreover, some interesting genes like BCL2, SNCA, CASP1 are selected and a significant overlap exists with the signature by Scherzer et al.

Discussion

The reliability of the statistical method is confirmed by the signature, that includes genes known to be involved in PD (like SNCA) and genes that are known to be implicated in other neurodegenerative diseases (like BCL2, CASP1). The signature is also intersecting with the one by Scherzer et al. (e.g. ST13, BCL11B). The functional characterization of the signature confirmed that the majority of the genes are included in categories already known to be affected by Parkinson (metabolism, cell related pathways as signaling and apoptosis, the nervous system and immune system related pathways).

Presenting Author

Margherita Squillario (margherita.squillario@unige.it)
Dipartimento di Informatica, Verona and DISI, Genova

Author Affiliations

(1) Dipartimento di Informatica, Universita degli Studi di Verona (2) Dipartimento di Informatica e Scienze dell'Informazione, Universita degli Studi di Genova

D-7. Feature elimination for one-class microRNA target prediction and gene identification

Khalifa W (1,2,), Yousef M (1,3)*

Different studies have described the use of two-class machine learning to predict microRNAs (miRNAs) gene target or gene identification. We present study using one-class machine learning for miRNA target discovery and gene identification. We compare one-class to two-class approaches using a Backward zero-norm for feature selection to Improve model performance, provide faster and more cost-effective models and gain a deeper insight into the underlying processes that generated the data.

Materials and Methods

we design duplex structure and sequence features in order to represent each example for the learning process. For each feature's vector we define a zero-norm to be the non-zero values over the given positive examples and chose different levels of threshold to determine a feature as irrelevant. We chose four one class methods to compare for miRNA discovery and gene identification: Support Vector Machines (SVM), Gaussian, Kmeans and K-Nearest Neighbor. Also, four two class methods: Random forest, Naïve Bayes, C4.5 and SVM. The WEKA software was used as implementation of the two-class classifier.

Results

Of all the one-class (OC)methods tested based on the all features, we found that most of them gave similar accuracy that range from 0.81 to 0.89 while the two-class gave 0.93-0.99 accuracy. the zero-norm feature selection with the OC improves the performance dramatically and makes the OC compete with two-class. It is clear from the accuracy plot of OC classifiers with different zero-norm thresholds that the performance is improving as the number of features is decreasing until a specific level. Interestingly, using zero-norm feature selection improves the results to reach accuracy of 0.96

Discussion

The current results show that it is possible to build up a classifier based only on positive examples yielding a reasonable performance. Moreover, using the zero-norm feature selection with the one-class approaches is essential to improve the performance to reach the two-class performance. Clearly the one-class is more sensitive to non-relevant features than the two-class. It is known that obtaining a biological data is an expensive process. The successful of the one-class classifier over a biological data can reduce the cost of generat-ing a such data by not requiring the control data.

Presenting Author

Waleed Khalifa (khwalid@hotmail.com)
The Institute of Applied Research- the Galilee Society

Author Affiliations

1.The Institute of Applied Research- the Galilee Society, P.O. Box 437 Shefa Amr,ZIP 20200, Israel 2. Computer Science, The College of Sakhnin, Sakhnin, ZIP30810, Israel. 3. Al-Qasemi Academic College, Baqa Algharbiya, ZIP 30100, Israel.

Acknowledgements

This project is funded in part under a grant with the ESHKOL scholarship program of the ministry of science culture and sport to support KW.

D-8. Computational identification of synonymous SNPs in the human genome and their potential role in disease

Wood L (1,2,), Ramsay M (2)*

The potential phenotypic effects of synonymous SNPs (sSNPs) have long been overlooked. Although several sSNPs are no longer thought to be silent no one has identified which sSNPs may contribute to disease phenotypes on a genome-wide scale, nor has a tool been developed that may identify a functional role for sSNPs in disease. Using available bioinformatics tools we have predicted the potential functional impact of many sSNPs within the human genome, and shown that more sSNPs could contribute to phenotypic variation than originally thought. The analysis tool will aid in the discovery of potential phenotypic effects of sSNPs within genes associated with disease.

Materials and Methods

Galaxy was used to retrieve all synonymous SNPs (sSNPs) from dbSNP130. Python scripts were written to calculate the change in codon-usage frequency caused by each sSNP. sSNPs that could potentially cause a change in mRNA secondary structure will be identified using RNAfold and RNAdist. ESEfinder 3.0 and MaxEntScan were used to determine which sSNPs could potentially result in aberrant splicing. BioMart was used to retrieve genes associated with disease from OMIM. sSNPs predicted to have a potential functional impact were then assessed to determine whether they were within genes previously associated with disease. MySQL will be used to create a searchable on-line database and analysis tool.

Results

Of the 72 504 synonymous SNPs (sSNPs) identified approximately 70% had no predicted change in function. The remaining sSNPs were predicted to have a potential functional impact by one or more of the molecular mechanisms. 35% of the sSNPs were predicted to cause a significant change in codon-usage frequency, whereas approximately 13% of the sSNPs were predicted to cause aberrant splicing. sSNPs that could potentially cause a change in mRNA secondary structure are currently being identified. 20% of the sSNPs were predicted to be within genes associated with disease, many of which were predicted to contribute to at least one of the molecular mechanisms.

Discussion

sSNPs that cause a change in codon-usage frequency or mRNA secondary structure may alter translational-kinetics and protein folding. In addition, sSNPs that disrupt splice-site consensus sequences may cause aberrant splicing, changing the protein product. Thus, a sSNP that contributes to any of these molecular mechanisms may cause a change in protein structure and function. Of the sSNPs published within dbSNP130, about 30% were predicted to have a potential functional impact, many of which were within genes associated with disease. Thus, more sSNPs may have a functional impact than originally thought and the potential role of sSNPs in disease should therefore not be underestimated or neglected.

Presenting Author

Lee-Ann Wood (123456.wood@gmail.com)
Wits Bioinformatics, University of the Witwatersrand, Johannesburg, South Africa

Author Affiliations

(1). Wits Bioinformatics, University of the Witwatersrand, Johannesburg, South Africa (2). Division of Human Genetics, National Health Laboratory Service, School of Pathology, University of the Witwatersrand, Johannesburg, South Africa

Acknowledgements

Cape Biotech and National Research Foundation

D-9. The potential functional role of chimeric transcripts in higher eukaryotes

Frenkel Morgenstern M (1,), Valencia A (1)*

Chimeric RNAs which are produced by trans-splicing of two or more distinct transcripts, have been reported in different organisms including fruit fly, mouse, and human [1,2]. Thousands of chimeric transcripts were identified among the variety of ESTs for these organisms. In theory, these transcripts can greatly contribute to the complexity of the transcriptome and proteome of organisms [1, 2], and nowadays are validated by the new generation sequencing technologies. However, a question of how the trans-splicing machinery if exists can be used to design new functional proteins remains unclear.

Materials and Methods

We have studied the potential functional role of the chimeric proteins in human and mouse. For this purpose, we annotated both of the two proteins participating in creation of the chimeric transcripts described by Li et al [2].

Results

We found that chimeras might integrate the transmembrane domains of some membrane proteins in their formation. Moreover, chimeras seem to contain full exons of the translated proteins, where the activation domain of one protein might be missing, but DNA binding domain remains in the resulting chimeric protein. In addition, for proteins, which function as dimers, the functional domains sometimes are missing, but dimerization domain is still present. Chimeras also integrate signal peptides from the proteins forming the chimera that can allow changes in its localization.

Discussion

To conclude, it seems that chimeric RNAs resulting from trans-splicing, if they are translated to proteins, could have a functional role suggested by our analysis, and may produce in some cases the dominant negative effect in the cells.

Presenting Author

Milana Frenkel-Morgenstern (mmorgenstern@cnio.es)
Spanish National Research Centre (CNIO)

Author Affiliations

(1) Spanish National Research Centre (CNIO), Madrid, Spain

Acknowledgements

The work of Milana Frenkel-Morgenstern is supported by the CNIO-Caja Navarra International Postdoctoral fellowships.
References: [1] Gingeras TR, Implications of chimeric non-co-linear transcripts, 2009, Nature, 461, pp 206-211. [2] Li X, Zhao L, Jiang H, Wang W, Short homologous sequences are strongly associated with the generation of chimeric RNAs in eukaryotes, 2009, J Mol Evol, 68:56-65.

D-10. A "Functional Signature" for analysing annotation space

del Pozo A (1) , Tress M (1), Valencia A (1)*

The hypotheses generated by homology-based functional annotation methods are formulated under the assumption that similar sequences have evolved from common ancestors and that functional roles are likely to have been conserved. Many studies have attempted to establish pairwise sequence identity cutoffs that guarantee homology. What is missing from these analyses is an exhaustive investigation of protein annotation space. Here we describe the current state of functional annotations in protein sequence databases with emphasis on those cases where reliable functional transfer should be possible

Materials and Methods

We selected a sample set of Uniref90 entries and obtained a subset of 6385 homologous clusters with at least one confident UniprotKB sequence. We used this subset to study the associated ’annotation space’ as snapshot of the more complex and generalized ’functional space’. We calculated a 'Functional Signature' for each cluster based on the quality of the annotations, the density of the local annotation space and the consensus of the functional terms (GO terms and EC numbers) with which the member sequences of each cluster are annotated.

Results

The 'Functional Signatures' generated for each cluster allowed us to define the annotation space of the clusters from the Uniref90 sample set. Over a third of the clusters were annotated with high quality functional terms and 20% of the clusters had at least one unannotated member, which indicated that homology-based functional annotation was possible in an important percentage of the clusters.

Discussion

Here we have introduced a new term, the 'Functional Signature' as a means of analysing the state of the annotation space of protein sequences. The analysis suggests that annotation space is unevenly populated with a large number of gaps that will need to be filled by non-homology based methods. For those clusters that are annotated with good quality functional annotations the functional information is not homogeneous, which shows that methods for homology-based functional transfer are still relevant.

Presenting Author

Angela del Pozo (adelpozo@cnio.es)
Structural Bioinformatics Group, Spanish National Cancer Research Centre (CNIO)

Author Affiliations

(1)Structural Bioinformatics Group, Spanish National Cancer Research Centre (CNIO). Madrid, Spain

Acknowledgements

ENFIN (Enabling System Biology). European Network of Excellence with contract number LSHG-CT-2005-518254 funded by the EU within its FP6 Programme.

D-11. Predicting small molecule ligand binding sites and catalytic residues with firestar/FireDB

López G (1,), Maietta P (1), del Pozo A (1), Rodríguez JM (2), Valencia A (1,2), Tress ML (1)*

Genome sequencing projects have lead to a surge in the number of protein sequences that lack experimental functional data. Experimental approaches for function characterization are expensive and difficult to automate and this has meant that researchers have turned increasingly to computational methods to try to close the gap between the number of new unannotated sequences and the number of sequences with known function. Frequently the most interesting functional information can be found at the level of the amino acid residues implicated in molecular interactions, catalysis or regulation.

Materials and Methods

The basic principle of the pipeline is the homology-based transfer of validated functional residues, supported by local sequence conservation. Previous versions of firestar required human interpretation of the results. Now the whole process has been automatized and a new web interface has been made available. Likely biological ligands have been separated from ligands present only as “contaminants”. Additionally our in house automatic pipeline is able to produce high quality results in high-throughput mode, which has allowed firestar to be incorporated into several prediction pipelines.

Results

We carried out testing of the server during the CASP8 experiment. We made predictions for the ligand binding prediction category, in which predictors have to predict residues in contact with small ligands. Although we were not allowed to compete officially in this category, testing showed that firestar would have been the best predictor in CASP8. The method was able to correctly detect 95% of the binding sites from the manually curated datasets. We have made improvements to the server accuracy for CASP9, in which firestar is currently participating and we hope to repeat the result from CASP8.

Discussion

We are currently using firestar to annotate functional residues for the human genome. The server is predicting functional residues for all annotated splice variants in order to evaluate whether functional residues are lost as a result of alternative splicing. In addition firestar has allowed us to analyse functional sites in protein coding genes, focusing on metal binding sites and cofactor binding sites. Finally we hope to extend reliable annotations from the Catalytic Site Atlas to all proteins in the human genome.

URL

http://firedb.bioinfo.cnio.es

Presenting Author

Michael L Tress (mtress@cnio.es)
Structural Biology and Biocomputing Programme, Spanish National Cancer Research Centre (CNIO)

Author Affiliations

(1) Structural Biology and Biocomputing Programme, Spanish National Cancer Research Centre (CNIO), C./ Melchor Fernandez Almagro, Madrid, Spain. (2) Spanish National Bioinformatics Institute (INB), Spanish National Cancer Research Centre (CNIO), C./ Melchor Fernandez Almagro, Madrid, Spain.

D-12. Classifying substrate specificities of membrane transporters from Arabidopsis thaliana

Schaadt NS (1,), Christoph J (1), Helms V (1)*

The transport of molecules across biological barriers such as lipid membranes is catalyzed by integral transmembrane proteins which act either as passive channels or as active carriers selective for a particular substrate or substrate group. Although the amino acid sequences of many putative membrane transporters are known, their specific functions remain often unknown due to the large experimental effort involved in determining substrate specificities. Therefore, it is highly desirable to develop computational methods that may aid in the substrate annotation of membrane transport proteins.

Materials and Methods

For all gene products from Arabidopsis thaliana that are annotated in the Aramemnon database as membrane transporters, we constructed subsets of those transporter sequences that are annotated to transport either amino acids, oligopeptides, phosphates, or hexoses. We ranked all transporters according to their similarity to the considered substrate class. The similarity was estimated by the Euclidean distance between the amino acid frequency of the transporters and the average value of the substrate class.

Results

Substrate classification based on the simple amino acid frequency had an accuracy of 75% or higher. Integrating higher-order sequence information, the frequencies of amino acid pairs, or from profile analysis further improved the prediction performance to 80-90%. For comparison, the accuracy for randomly compiled substrate classes was only around 30%.

Discussion

We developed a novel computational approach to predict substrate specificities of membrane transport proteins from their amino acid sequence. We found that membrane transporters belonging to different transporter families but transport the same substrate have a significantly higher similarity of their amino acid frequencies that random pairs of transporters. Thus, the amino acid characteristics of membrane transporters appear related to the physicochemical properties of the transported substrates.

Presenting Author

Nadine S. Schaadt (nschaadt@bioinformatik.uni-saarland.de)
Center for Bioinformatics, Saarland University

Author Affiliations

Center for Bioinformatics, Saarland University

D-13. Molecular modeling and comparison by docking and molecular dynamics of the interaction between fatty acids and proteins and EgFABP2 EgFABP1

Paulino Z-M (1,), Esteves A (2)*

The fatty acid binding proteins (FABPs) are cytosolic proteins of low molecular weight (14-15 kDa) and very abundant. They are able to bind fatty acids and other small hydrophobic molecules. Several types of tissue-specific FABPs have been identified in vertebrates and more recently, in two FABPS - EgFABP1 and EgFABP2 - in the flatworm parasite Echinococcus granulosus. Both proteins have a similarity of 96% at amino acid level and a significant sequence similarity with mammalian members of the subfamily of cardiac FABPs.

Materials and Methods

In determining whether there are functional differences between the two parasite proteins, and considering that the functional specificity could be linked to differential ligand affinities, we have initiated the in silico analysis of the binding energies of each protein by different ligands (arachidonic, palmitic, oleic and linoleic acids), using the software package MOE 2007 -2009.

Results

The energetic docking results (scores) were not conclusive, giving similar interaction energies for all ligands. Then, the optimal energy structures were taken as the starting point of 4 nanoseconds molecular dynamics procedures in which the solvent was included explicitally and a full flexibility allowed. After MD simulation, all results were averaged and the standard deviations evaluated. All results showed low standard deviations, being the differences between different ligands, significants.

Discussion

EgFABP1 showed statistically better interaction energy with arachidonic acid by the other ligands, while EgFABP2 showed better interaction energy for oleic acid. Three main contacts of carboxylic head of all acids were made with side chains of the conserved residues Tyr 129 Arg 107 and Arg 127. Variations between different affinities are given through the effect of hydrogen bonding of those residues as well as a water cluster around the polar head of ligands. This observation evidencenced the importance of the influence of solvent in the network of contacts to strengthen the interaction.

URL

http://www.fq.edu.uy

Presenting Author

Margot Paulino Zunini (margot@fq.edu.uy)
Facultad de Química

Author Affiliations

1. LaBioFarMol, DETEMA, Facultad de Química, UdelaR, General Flores 2124, 11600, Montevideo, Uruguay 2. Sección Bioquímica, Instituto de Biología, Facultad de Ciencias UdelaR, Iguá 4225, Montevideo, Uruguay

Acknowledgements

Agencia National de Investigación e Innovación (ANII). Uruguay. Programa de Desarrollo de Ciencias Básicas (PEDECIBA), Uruguay.

D-14. GreenPhylDB version 2: web resources for comparative and functional genomics in plants

Rouard M (1,), Guignon V (1,2), Aluome C (1,2), Laporte MA(2), Droc G(2), Walde C(1), Zmasek CM(3), Perin C(2), Conte MG(1)*

With the increasing number of plant genomes being sequenced, a major objective is to transfer annotation between genome models and other species of interest. Orthology inference is one of the most reliable strategies for functional annotation

Materials and Methods

GreenPhylDB contains gene families being annotated, computational analyzes and external references (e.g. InterPro, KEGG, Swiss-Prot, Pubmed) related to all sequences. Once manually annotated (i.e. properly named and classified), gene families are finally processed by phylogenetic analyses to distinguish orthologous and paralogous gene.

Results

The website offers a range of user-friendly tools to query the data. These resources will be particularly helpful to molecular biologist for gene discovery and gene function inference.

Discussion

A better understanding of genome evolution will contribute to elucidate the molecular basis of important agronomic traits and therefore facilitate ongoing plant breeding efforts.

URL

http://greenphyl.cirad.fr

Presenting Author

Mathieu Rouard (m.rouard@cgiar.org)
Bioversity International

Author Affiliations

1 Bioversity International - CfL programme Parc Scientifique Agropolis II, 34397 Montpellier, France 2 CIRAD, Department BIOS, UMR DAP - TA40/03, 34398 Montpellier, France 3 Sanford-Burnham Medical Research Institute, 10901 North Torrey Pines Road, La Jolla, CA 92037, USA.

Acknowledgements

Generation Challenge Programme (GCP)

D-15. Collaboration-based function prediction in protein-protein interaction networks

Rahmani H (1,), Blockeel H (1,2), Bender A (3)*

Existing techniques assume that proteins that are topologically close in the network tend to have similar functions. We hypothesize that better predictive accuracy can be obtained by generalizing this assumption. We call two functions collaborative if proteins with one function often interact with proteins performing the other function. Our hypothesis is that techniques that extract such function collaboration information from networks, and exploit it, can yield better predictions. We propose and evaluate two such techniques.

Materials and Methods

We propose two methods implementing the function collaboration hypothesis. In the first method, first we extract the collaborative function pairs from the whole network. Then, in order to make a prediction for an unclassified protein, we extract the candidate functions based on the position of the protein in the network. Finally, we calculate the score of each candidate function. High score candidate functions are those which collaborate more with the neighborhood of unclassified protein. The second method adopts Self Organizing Map (SOM) for modeling the function collaboration in PPI networks

Results

We selected two methods, Majority Rule and Functional Clustering, as representatives of the similarity based approaches. We compared our collaboration based methods with these similarity based methods on three interaction datasets: Krogan, DIP-Core and VonMering. We examined up to five different function levels and we found classification performance according to F-measure values indeed improved, sometimes by up to 17 percent, over the benchmark methods employed.

Discussion

To our knowledge, this is the first study that considers function collaboration for the task of function prediction in PPI networks. We view biological process as an aggregation of individual protein functions and our hypothesis is that topologically close proteins have collaborative functions. We propose two methods that predict protein functions based on function collaboration. We perform a comprehensive set of experiments that reveal a significant improvement (ranging from 3% to 17%) of F-measure values compared to existing methods.

Presenting Author

Hossein Rahmani (hrahmani@liacs.nl)
Leiden Institute of Advanced Computer Science

Author Affiliations

(1)Leiden Institute of Advanced Computer Science, Universiteit Leiden Niels Bohrweg 1, 2333 CA Leiden, The Netherlands (2) Department of Computer Science, Katholieke Universiteit Leuven, Celestijnenlaan 200A, 3001 Leuven, Belgium (3) Division of Medicinal Chemistry, Leiden/Amsterdam Center for Drug Research, Universiteit Leiden, 2333 CC Leiden, The Netherlands

Acknowledgements

This research is funded by the Dutch Science Foundation (NWO) through a VIDI grant. Andreas Bender was funded by the Dutch Top Institute Pharma, project number: D1-105.

D-16. SVM feature selection applied to lignocellulose degradation

Trukhina Y (1,), McHardy A (1)*

The study of the lignocellulose degradation process is an area of substantial scientific interest, not only because of the general scientific importance of this topic, but also because of the possible commercial and ecological benefits from controllable biodegradation of cellulosic waste.

Materials and Methods

In this work we investigate this problem, and more generally, phenotype prediction from genomic data, which is one of the most important tasks in computational genomics, with an approach based on support vector machine (SVM) models. More specifically, we apply feature selection with an L1-regularized linear SVM to our data set, which we created based on information provided by the IMG/M database.

Results

Our approach results in a good phenotype prediction accuracy and a meaningful set of selected features.

Discussion

Our results are consistent with the current knowledge of microbial cellulose degradation.

Presenting Author

Yulia Trukhina (trukhina@mpi-inf.mpg.de)
Computational Genomics and Epidemiology Group, Max Planck Institute for Computer Science, Saarbruecken, Germany

Author Affiliations

Computational Genomics and Epidemiology Group, Max Planck Institute for Computer Science, Campus E1 4, 66123 Saarbruecken, Germany

Acknowledgements

Max-Planck Society

D-17. Substrate-specific HMMs for the classification of adenylation and acyltransferase domains of NRPS/PKS systems

Khayatt B I (1,), Siezen R J (1,2,3), Francke C (1,3)*

NRPS/PKS systems produce bioactive peptides and polyketides in a variety of microbes and plants. The molecules range from antibiotics to kill competitors (e.g. penicillin and erythromycin) to surfactants produced to escape a biofilm environment (e.g. surfactin). The ability to identify substrate specificity of the adenylation (A) and acyltransferase (AT) domains will be very helpful in the rapid discovery of new products of these systems.

Materials and Methods

NCBI, UniProt and NRPS/PKS-specialized databases were used to retrieve A and AT sequences of experimentally verified NRPSs and PKSs. The retrieved sequences were aligned and the residues relating to substrate specificity were extracted. Using these residues, substrate-specific Hidden Markov Models were generated and applied and ‘the recognition power’ of these different A and AT substrate models was evaluated.

Results

We collected a set of 469 A and 164 AT domain sequences and determined the substrate-specific residues by splitting the set on basis of substrate specificity after alignment. We found that there was more variation in the AT domains. An initial classification procedure using Neighbor Joining trees, as applied before by others, appeared to limit the power of recognition to around 85%. We therefore created substrate-specific HMMs and tested whether these would improve performance. Indeed, using these models, the recognition power improved to 95% and 99% for A and AT domains, respectively.

Discussion

We found that the substrate specificity of the A and AT domains is largely captured in a specific set of domain residues and highly specific HMMs could be generated. The methodology to generate substrate-specific HMMs will be helpful in detecting substrate specificity of the sub-classes of other enzymes that have so far no characterized structural models as references for residue numbering.

Presenting Author

Barzan I. Khayatt (b.khayatt@cmbi.ru.nl)
Radboud University Nijmegen Medical Centre, NCMLS, CMBI

Author Affiliations

1) Center for Molecular and Biomolecular Informatics (CMBI) (260), NCMLS, Radboud University Nijmegen Medical Center, PO Box 9101, 6500HB Nijmegen, The Netherlands 2) NIZO food research, PO Box 20, 6710BA Ede, The Netherlands 3) TI Food and Nutrition, Kluyver Centre for Genomics of Industrial Fermentation and Netherlands Bioinformatics Centre, PO Box 557, 6700AN Wageningen, The Netherlands

D-18. Stochastic modeling of proteolytic activity from LC-MS data

*Dittwald P (1, *), Gambin A (1), Ostrowski J (2) , Karczmarski J (2)*

The motivation for model is a hypothesis that the activity of specific enzymes present in the serum samples of patients and healthy donnors differ significantly giving the possibility to distinguish colorectal cancer and healthy samples.

Materials and Methods

Our approach integrate the existing knowledge about proteases' activity stored in MEROPS database with the efficient procedure for estimation the model parameters. The proteolytic activity is modelled with the use of Chemical Master Equation. Assuming the stationarity of the Markov process we calculate the expected values of digested peptides in the model. The parameters are fitted to minimize the discrepancy between those expected amounts and the peptide intensities observed in the tandem mass spectrometry data. Constrained optimization problem is solved by Levenberg-Marquadt algorithm.

Results

We tested our approach on 39 colorectal cancer and healthy control subjects. The set of proteolytic enzymes probably involved in the disease process has been identified. Using these enzymatic activities occur to be meaningful in discriminating between healthy and colorectal cancer samples. Our preliminary resuts show the efficiency and accuracy of the estimation procedure.

Discussion

The model significantly extends those proposed recently in [Klu08]. In the current approach the parameters of the model inferred from LC-MS/MS correspond directly to the activity of specific enzymes present in the serum samples of patients and healthy control subjects. The scanning of biological literature show that many of enzymes identified by our approach are regarded to be involved in colorectal cancer process. Bibliography: [Klu08] Kluge, B., Gambin, A. & Niemiro, W., Modeling Exopeptidase Activity from LC-MS Data (2009), Journal of Computational Biology, Vol.16, No.2, Pp.395-406

URL

http://bioputer.mimuw.edu.pl/papers/proteolysis/

Presenting Author

Piotr Dittwald (piotr.dittwald@students.mimuw.edu.pl)
Institute of Informatics, University of Warsaw

Author Affiliations

1: Institute of Informatics, University of Warsaw, Banacha 2, 02-097 Warsaw, Poland 2: Department of Gastroenterology and Hepatology, Medical Center for Postgraduate Education at the Maria Sklodowska-Curie Memorial Cancer Center, Institute of Oncology, 02-781 Warsaw, Poland

Acknowledgements

This research is supported in part by Polish Ministry of Science and Educations grants N301 065236 and N206 356036.

D-19. Mining posttranscriptional regulatory features by comparing total and polysomal mRNA profilings

Re A (1,), Tebaldi T (1), Segata N (2), Passerini A (2), Blanzieri E (2), Quattrone A (1)*

Posttranscriptional events such as export, transcript stability and translational regulation contribute substantially to the final protein readout. They underlie the disparity recurrently noted between the steady-state levels of the mRNAs and those of the proteins. Little is known on how often each mRNA is sensitive to posttranscriptional activities and how these activities influence mRNAs. The aim of this study is to quantify the extent of posttranscriptional control and to identify the relevant features, from the general ones providing aspecific control to those regulating specific mRNAs.

Materials and Methods

We combined total and polysome-bound RNA profilings in eight conditions in order to systematically estimate by what degree the total and polysomal fractions of an mRNA differ. As the mRNAs actively engaged in polysomes integrate all the events contributing to the rate of protein synthesis, they have been used as a proxy for newly synthesized proteins. We subsequently computed to what extent general and mRNA-specific, structural and sequential features of 5' and 3' UTRs of expressed mRNAs enable to characterize classes of mRNAs with different translational efficiency estimated as above.

Results

From a systems-level point of view quantifying the mRNA engagement with posttranscriptional control uncovers distinct populations. A general trend among conditions seems to indicate a higher informativity of 3'UTR features with respect to 5'UTR ones. This trend was especially evident for aspecific features like UTRs length, folding propensity and GC-content. Furthermore mRNA-specific sequential and structural features have been found to be required for characterizing the mRNA populations. The specific trans-acting factors recognizing the cis-acting elements have been partially determined.

Discussion

The combination of total and polysomal mRNA profiling uncovers a widespread uncoupling between transcriptome and translatome. General 5' UTR features are less important than suggested to explain the existence of weakly translated mRNAs. mRNA specific features should be taken into account and the responsible corresponding trans-acting factors identified, highlighting the relative contribution of coding and non-coding factors. A better characterization of post-transcriptional regulation would also require a more complex modeling of the interaction among multiple regulatory elements.

Presenting Author

Angela Re (re@science.unitn.it)
Centre for Integrative Biology, University of Trento

Author Affiliations

(1) Centre for Integrative Biology, University of Trento, Trento, Italy (2) Department of Information Engineering and Computer Science, University of Trento, Trento, Italy

D-20. Prototypes of elementary functional loops unravel evolutionary connections between protein functions

Goncearenco A (1,2,), Berezovsky IN (1)*

Earlier studies of protein structure revealed closed loops with a characteristic size 25-30 residues and ring-like shape as a basic universal structural element of globular proteins. Elementary functional loops (EFL) have specific signatures and provide functional residues important for binding/activation and principal chemical transformation steps of the enzymatic reaction. The goal of this work is to show how these functional loops evolved from pre-domain peptides, and to find a set of prototypes from which the EFLs of contemporary proteins originated.

Materials and Methods

We describe a computational method for deriving prototypes of elementary functional loops based on the sequences of complete genomes. The procedure comprises the iterative derivation of sequence profiles followed by the hierarchical clustering of profiles. We propose a scoring function that weights profile positions proportional to the information content on position, allowing to discriminate between matches that carry a specific signature from non-specific ones. The statistical significance of the scores is calculated from the empirical distribution of the reshuffled profile scores.

Results

We developed a computational procedure for deriving prototypes of EFLs, obtained prototypes from the set of archaeal proteomes, considered several prototypes in detail, delineated connections between domains superfamilies using the most abundant prototypes, and exemplified how combinations of EFLs result in specific enzymatic function.

Discussion

The existence of EFLs in different folds and functions makes it possible to survey subtle evolutionary relations, originating from the pre-domain evolution of protein structure. It suggests that contemporary enzymatic functions are constructs of different sets and combinations of elementary chemical functions. Obtaining the full collection of prototypes with elementary functions will make it possible to (a) predict enzymatic functions via determining EFLs corresponding to prototypes; (b) (re)design folds with desired functions by building constructs from necessary elementary functional loops.

Presenting Author

Alexander Goncearenco (alexandr.goncearenco@uni.no)
University of Bergen, Bergen Center for Computational Science

Author Affiliations

(1) Computational Biology Unit, Bergen Center for Computational Science; (2) Department of Informatics, University of Bergen, N-5008 Norway.

Acknowledgements

FUGE-II Norwegian functional genomics platform

D-21. Utilization of proteins and nucleic acids in the study of gene function: a comparative review

Mwololo JK (1), Karaya HG(2), Munyua JK(3), Muturi PW(1), Munyiri SW(1)*

Proteomics is one of the fastest growing areas in areas of research, largely because the global-scale analysis of proteins is expected to yield more direct understanding of function and regulation than analysis of genes. Protein structure characterizes its function and a protein sequence that relates to a known structure forms a basis for identifying gene function. Proteins are encoded by the genome (genes), and the set of proteins encoded by the genome, including the added variation of post-translational modification, constitute the proteome.

Materials and Methods

The major role of protein structure is to characterize function. Proteins are encoded by the genome and are involved in cellular metabolic activities. DNA can provide more information than proteins due to the degeneracy of the genetic code and the presence of large non-coding stretches.The proteins are dynamic and interacting molecules, and their state of instability can make proteomic snapshots difficult. Use of DNA to understand such complexity is not adequate because problems of redundancy, pseudogenes and sequencing create inaccuracies. Measuring the intermediate step between genes and proteins bridges the gap between the genetic code and the functional molecules that run cells.

Results

It is clear that proteins are more amenable to prediction of gene function and therefore proteomics will increase in importance since analyses of proteins is expected to yield more direct understanding of function and regulation of genes than would be achieved through analysis of genes themselves. However, there is need to integrate genomics, transcriptomics and proteomics to facilitate understanding of normal cell development, function and response to diseases.

Discussion

The challenges encountered in identifying the biochemical and cellular functions of the many gene products which are currently not characterized has necessitated the use of the proteome. The development of automated methods for the annotation of predicted gene products with functional categories is becoming important. Compared to the study of the genetic code proteomics may allow greater understanding of the complexity of life and the process of evolution due to the large number of proteins that can be produced by an individual organism. A major challenge to proteomics is that proteins are dynamic and interacting molecules and their variability can complicate detailed studies on gene function.

Presenting Author

James K. Mwololo (mwololojames@yahoo.com)
MAKERERE UNIVERSITY

Author Affiliations

1.Makerere University, Faculty of Agriculture, Crop Science Department, P.O. Box 7062 Kampala, Uganda; 2.International Maize and Wheat Improvement Centre (CIMMYT), P.O. Box 1041-00621, Nairobi, Kenya; 3.University of Nairobi, P.O. Box 30197-00100 Nairobi, Kenya.

Acknowledgements

Support from my co-authors in preparation of this manuscript.

D-22. ClanTox: A predictor tool for toxin-like proteins reveals 500 such proteins within viral genomes

Naamati G(1,), Askenazi M(2,3), Linial M(2)*

Animal toxins operate by binding to receptors and ion channels. These proteins are short and vary in sequence, structure and function. Sporadic discoveries have also revealed endogenous toxin-like proteins in non-venomous organisms. Viral proteins are the largest group of quickly evolving proteomes. We tested the hypothesis that toxin-like proteins exist in viruses and that they act to modulate functions of their hosts.

Materials and Methods

Data collection: We collected viral proteins from UniProt. We started with 773,000 sequences which were reduced to 26,000. Vector construction: Each protein sequence was represented as a vector of 545 sequence derived numerical features. Among these features we included amino acid frequencies, length, and specific amino acids and their spacing, with special consideration for Cysteine. Training sets: We used a training set of 924 sequences listed as ionic channel inhibitors. Learning algorithm: We used a meta classifier based on a boosted stump algorithm (Adaboost).

Results

We updated and improved a classifier for proteins resembling short animal toxins that is based on a machine learning method. We applied identify toxin-like proteins among short viral proteins. Among the ~26,000 representatives of such short proteins, 510 sequences were positively identified. We focused on the 19 highest scoring proteins. Our predictor was shown to enhance annotation inference for many ‘uncharacterized’ proteins. We conclude that our protocol can expose toxin-like proteins in unexplored niches and enhance the systematic discovery of novel cell-modulators for drug development.

Discussion

Clantox identifies short proteins that function as cell modulators that are beyond the function of toxins. Among the ~500 viral positive predictions we identified a surprising number of conotoxin-like proteins. Finding of a large number of conotoxin-like peptides in viruses is intriguing. Two scenarios may account for this observation: (i) fast evolving convergent evolution of short secreted proteins(ii) Genetic material exchange from hosts to viruses. We postulate that some of the sequences identified in this study carry a potential for drug development.

URL

http://www.clantox.cs.huji.ac.il/

Presenting Author

Guy Naamati (guy.naamati@mail.huji.ac.il)
The Hebrew University of Jerusalem

Author Affiliations

1) Department of Computer Science, and Engineering, The Hebrew University of Jerusalem, Israel. 2) Department of Biological Chemistry, The Hebrew University of Jerusalem, Israel. 3) Blais Proteomics Center, Dana-Farber Cancer Institute, Boston, MA, USA.

Acknowledgements

This study is supported by the grant agency of the ISF and BSF. G.N is supported by the SCCB, the Sudarsky Center for Computational Biology.

D-23. I-Patch web service: inter-protein contact prediction using local network information

Hamer R (1,2), Luo Q (2,3), Armitage J (1), Reinert G (1,2), Deane C (1,2), Krawczyk K (1,2,)*

Prediction of inter-protein contact sites and of specificity-determining residues is of utmost importance for the pharmaceutical industry. Since the set of specificity-determining residues should be contained within the set of all contact sites, identifying this set would be a step forward towards identifying the specificity-determining contact-sites. I-Patch achieves the goal of identifying contact-sites with 59% precision and 20% recall - better than the other existing methods.

Materials and Methods

Given input of 2 PDB files (proteins) and MSA's of their homologs (ideally with indication of known inter-homolog complexes), the propensity score for each residue is calculated. The algorithm relies on propensity data for a single residue, pair of residues or a triple of residues to be a contact site. The data was derived from a non-redundant set of 677 protein complexes and 1150 proteins with two domains on a single chain, yielding a set of 1827 distinct proteins. Fitting and blind data sets consisted of 31 proteins each - sets were non-overlapping with those used to calculate propensities.

Results

I-Patch achieves 59% precision with 20% recall on the non-redundant blind data set of 31 proteins. It performs better than other existing methods (ELSC, SCA, OMES, MI and McBASC). The best results from amongst other algorithms on this data set are by EBMcBASC - 36% precision with 20% recall. I-Patch predictions on proteins involved in bacterial chemotaxis correctly identified 13 out of 36 contact sites. 27 predicted contact sites do not correspond to any known contact sites, some of them however, seem to lie on inter-domain interfaces or previously unknown interfaces.

Discussion

Method for identifying protein contact sites was developed by treating the protein complex as a network of possible contact sites. The algorithm performs better than other existing methods on the same data set and seems to be identifying previously unknown interaction interfaces. The program is currently available for MATLAB from the project website but a web-based service will be available by the end of August/early September (information about which will be available on the project website).

URL

http://www.stats.ox.ac.uk/research/bioinfo/resources

Presenting Author

Konrad Krawczyk (konrad.krawczyk@dtc.ox.ac.uk)
University of Oxford, Lady Margaret Hall

Author Affiliations

(1) Oxford Centre for Integrative Systems Biology, University of Oxford, Oxford, UK (2) Department of Statistics, University of Oxford, Oxford, UK (3) Department of Mathematics and Systems Science, National University of Defense Technology, Changsha, Hunan, China

D-24. Predicting candidate genes for biomass traits

Hassani-Pak K (1,), Kuo SC (1), Hanley S (1), Rawlings C (1)*

The phenotype-genotype problem starts with phenotypic variation and tries to determine which genes are involved. Quantitative trait loci (QTL) are chromosomal regions identified through linkage analysis that assign variation observed in a phenotype to a region on a genetic map. Even well-defined QTL, however, may encompass many potential candidate genes, perhaps hundreds. It is therefore hard to objectively choose underlying candidate(s) that drive the phenotype.

Materials and Methods

We are developing bioinformatics tools to support systematic analysis of QTL regions and prioritise genes for experimental analyses. Prioritisation is generally based on evidence that supports the role of a gene product in the biological process being investigated. The two most important bodies of information providing such evidence are bioinformatics databases and the scientific literature.

Results

In this work, we first present a knowledge base (KB) for the Poplar genome created by using comparative genomics, data integration and text mining methods of the freely available Ondex system (www.ondex.org). Poplar proteins in this KB are enriched with functional, phenotype and literature information. Second, we demonstrate our novel web-based tool to query the Poplar KB and analyse a QTL region to identify and prioritise candidate genes for complex phenotypes such us biomass production.

Discussion

We present a generic workflow that can be used for the functional annotation of any newly sequenced genome. Scientists working on functional genomics or population genetics, searching for new ways to explore their data and predicting candidate genes, will be interested in our novel visualization tool which enables users to examine QTLs and explore the underlying networks.

Presenting Author

Keywan Hassani-Pak (keywan.hassani-pak@bbsrc.ac.uk)
Rothamsted Research

Author Affiliations

Rothamsted Research, UK

Acknowledgements

BBSRC

D-25. Exploring residue interaction networks to understand protein structure-function relationships

Doncheva N-T (1,), Domingues F-S (1), Albrecht M (1)*

Studying the properties of individual amino acid residues and their pairwise interactions improves the understanding of protein structure-function relationships. Recent work showed that 2D residue interaction networks (RINs) derived from 3D protein structures provide new insights into the role of biologically important residues for protein function. In particular, RINs are very suitable for characterizing the effect of residue mutations on protein structure and function, for instance, if the mutated residue is not located close to the affected protein site.

Materials and Methods

A RIN is represented as a network in which the interacting nodes are the amino acid residues and the interaction edges represent their atomic contacts. We also use a more sophisticated method for RIN generation that considers non-covalent interactions between main and side chains of the amino acids and quantifies the strength of these interactions. For most protein structures deposited in the Protein Data Bank, we have generated precomputed RINs for download.

Results

We developed the Cytoscape plugin RINalyzer, which uses UCSF Chimera to support simultaneous viewing and exploring of the 2D network of residue interactions and the corresponding 3D protein structure. Among other visual analytics tools, RINalyzer offers the computation and visualization of a comprehensive set of centrality measures for discovering critical residues in the protein structure. Additionally, comparisons between RINs can be performed by constructing a combined network that allows for investigating the residue interaction differences between aligned 3D protein structures.

Discussion

Due to its rich functionality and user-friendly interface, RINalyzer is a versatile tool that facilitates the usage of RINs for investigating and understanding complex protein structure-function relationships.

Presenting Author

Nadezhda T. Doncheva (doncheva@mpi-inf.mpg.de)
Department of Computational Biology and Applied Algorithmics, Max Planck Institute for Informatics

Author Affiliations

(1) Department of Computational Biology and Applied Algorithmics, Max Planck Institute for Informatics, Germany

D-26. BioXSD: the canonical XML-Schema data model for bioinformatics web services

Kalaš M (1,2,), Puntervoll P (1), Joseph A (3), Bartaševičiūtė E (4), Töpfer A (1,5), Ison J (6), Blanchet C (3), Rapacki K (4), Jonassen I (1,2)*

The wide community of life-scientific groups offers a huge amount of public bioinformatics data sources and tools. These resources cover diverse areas of bioinformatics and follow diverse sets of implementation designs. More and more of the resources provide a standardised programmatic Web-service interface. However, we are witnessing a burden to smooth interoperability among the services: a lack of standard input and output data formats.

Materials and Methods

BioXSD has been developed by EMBRACE project partners, by analysing existing approaches, tools, data formats, ontologies, and the requirements of bioinformaticians and bench biologists. We defined the canonical exchange formats in a pure XML Schema, semantically annotated using SAWSDL standard by the EMBRACE Data And Methods (EDAM) ontology. EDAM defines formal semantics of the syntactic BioXSD data objects. Maintenance and further development of BioXSD will be done in an open collaboration with the world-wide bioinformatics community, curated by the BioXSD consortium.

Results

BioXSD is a candidate for standard, canonical exchange format for basic bioinformatics data types. BioXSD defines syntax for biological sequences, sequence annotations, alignments, and references to resources (data, databases, taxonomies, ontologies). We have adapted a set of Web services to use BioXSD as the input and output format, and implemented a test-case workflow. This demonstrates that the approach is feasible and provides smooth interoperability.

Discussion

Providers of bioinformatics databases and tools are encouraged to include BioXSD among the other supported input and output formats. BioXSD types can be further included in custom types, extended, or restricted, or can be used as an intermediate exchange format. We recommend the specialised standard XSDs (SBML, PDBML, MAGE-ML, PSI MI MIF, phyloXML, ...) to be used whenever applicable. BioXSD should be used for the most common, basic bioinformatics data-types. These basic data types were not previously standardised in a sufficiently interoperable way applicable to the world-wide Web services.

URL

http://bioxsd.org

Presenting Author

Matus Kalas (matus.kalas@bccs.uib.no)
Computational Biology Unit, Bergen Center for Computational Science

Author Affiliations

(1) Computational Biology Unit, Bergen Center for Computational Science, Uni Research, Bergen, Norway. (2) Department of Informatics, University of Bergen, Bergen, Norway. (3) Institut de Biologie et Chimie des Protéines, Centre National de la Recherche Scientifique & Université Claude Bernard Lyon 1, Lyon, France. (4) Center for Biological Sequence Analysis, Department of Systems Biology, Technical University of Denmark, Kongens Lyngby, Denmark. (5) Institute for Bioinformatics, Center for Biotechnology, Bielefeld University, Bielefeld, Germany. (6) European Bioinformatics Institute, EMBL, Wellcome Trust Genome Campus, Hinxton, Cambridge, UK.

Acknowledgements

This work was supported by the European Commission within FP6 [LHSG-CT-2004-512092, the EMBRACE project]; the Research Council of Norway within eVITA [178885/V30, eSysbio project, to M.K., P.P., A.T., I.J.]; l'Agence Nationale de la Recherche [ANR-06-CIS6-005, HIPCAL project, to A.J.]; and the Villum Foundation [grant for the Center of Disease Systems Biology, to E.B. and K.R.].

AUTHOR INDEX