LATE-BREAKING POSTERS

A-63. An Ensembl-based pipeline for microRNA prediction and expression profiling using Next Generation Sequencing data

*James N (1, *), Donepudi M (1), Spooner W (1), Watson M (2)*

Predicting miRNA loci and profiling miRNA expression based on short read sequences generated from small RNA libraries.

Materials and Methods

Our pipeline runs on the Ensembl eHive distributed processing system for which we have built wrappers for a number of best-in-class, open-source miRNA analysis software including RNAfold, MiPred, miRDeep and DroshaSVM.Ensembl databases are used for data storage, automatically integrating results with the latest genome annotations and providing an excellent and widely used interface for data access.

Results

We have developed a robust and highly scalable system for miRNA prediction from next generation sequencing data. It incorporates a number of open source miRNA prediction software, to which more can easily be added. The workflow system is compatible with several cluster architectures, including Sun Grid Engine, Condor, Platform LSF, Amazon Web Services or standalone.

Discussion

We have developed a robust and highly scalable system for miRNA prediction from next generation sequencing data. It incorporates a number of open source miRNA prediction software, to which more can easily be added. The workflow system is compatible with several cluster architectures, including Sun Grid Engine, Condor, Platform LSF, Amazon Web Services or standalone.

Presenting author

Nick P James (nick@eaglegenomics.com)
Eagle Genomics Ltd

Author Affiliations

(1) Eagle Genomics Ltd, Babraham Research Campus, Cambridge CB22 3AT (2) Bioinformatics Group, Institute for Animal Health (IAH), Compton, Newbury, RG20 7NN, UK

URL

`http://www.eaglegenomics.com`

A-64. The Goby file formats: towards scalable next-generation sequencing data analysis

Chambwe N (1,2), Dorff KC (1), Srdanovic M (1), Deng M (1), Andrews SJD (1,2) and Campagne F (1,2,)*

Next-Generation Sequencing (NGS) technologies continue to evolve rapidly, generating massive amounts of short-read sequence data. The sheer volume of NGS data introduces computational challenges in its processing and storage. NGS projects often require the ability to store sequence reads, alignments, base-resolution histograms, and to record specific subsets of reads. Current file formats support each of these types of information, but files encoded in widely used formats can become very large, creating scalability issues in most NGS analysis pipelines.

Materials and Methods

We used Gzip compression with the open-source Google Protocol Buffers library (http://code.google.com/p/protobuf/). Protocol buffers are advantageous because they support multiple languages (i.e., Python, Java, C++ and others), are cross-platform, flexible, and extensible. For instance, they allow older versions of software to read newer versions of a file format (forward compatibility) or new versions of the software to read older versions of a file format (backward compatibility).

Results

We have implemented compact and scalable file formats in the Goby software framework. Goby files are organized as chunks and support semi-random access in a very large file, which is useful to retrieve only slices of a read or alignment file for processing on a compute cluster. Goby file formats are also precisely specified with the Protocol Buffer schemas and are often significantly smaller than current standards. For instance, we find that Goby alignment files can be 2 to 5 times smaller than an equivalent alignment encoded in BAM format (compressed binary version of a SAM file).

Discussion

We have tested the Goby file formats by constructing an RNA-Seq analysis pipeline. This pipeline supports multiple aligners, including the Burrows Wheeler Aligner (BWA) and the Last aligner and has been tested using data from the Sequencing Quality Control (SEQC) (http://www.fda.gov/MAQC/SEQC), from four major sequencing platforms. The poster will discuss how the approaches implemented in Goby compare to the most widely used NGS file formats (Fasta/Fastq Elan text format, MAQ, SAM/BAM).

Presenting author

Fabien Campagne (fac2003@med.cornell.edu)
Weill Medical College of Cornell University

Author Affiliations

1 The HRH Prince Alwaleed Bin Talal Bin Abdulaziz Alsaud Institute for Computational Biomedicine; Weill Medical College of Cornell University, New York, NY 10021; 2 Department of Physiology and Biophysics, Weill Medical College of Cornell University, New York, NY 10021

URL

`http://campagnelab.org`

A-65. Quantitative transcriptome analysis using de novo assembly of RNA-seq from the non-model C4 species Cleome gynandra

Salmon-Divon M (1,), Aubry S (2), Rutherford K-M (2), Kelly K-A (2), Hibberd J-M (2), Bertone P (1)*

Due to the rapid development in NGS technologies, it is now feasible to perform quantitative analyses of transcriptomes without prior knowledge of an organism's genome. Here we describe the de novo assembly of Illumina-based RNA sequencing data to establish the Cleome gynandra transcriptome, and to assess global changes in gene expression during leaf development. We were able to identify and quantify accumulation of an estimated 60% of the leaf transcriptome, and demonstrate the unique possibilities opened by RNA-seq for rapid and cost-efficient transcriptome analysis of non-model organisms.

Materials and Methods

Using the Illumina Genome Analyzer platform, approximately 115 million sequence reads, 36 to 42 bp in length, were generated from six sequencing runs representing triplicates of young and mature leaves. We compared the de novo assemblies generated by two methods, Velvet/Oases and ABySS. We then assessed the accuracy of each transcriptome assembly by combining Illumina-based data with longer sequencing reads obtained by 454 pyrosequencing.

Results

A total of 118,611 contigs were produced in the assembly phase. BLASTX analysis of each contig against Arabidopsis proteome identified 15,215 genes, 1,692 of which were found to be differentially expressed between early and mature leaves. Gene ontology analysis revealed that major transcriptome reprogramming occurs in the developing leaves, where up-regulation of cell wall and protein synthesis genes is evident in earlier developmental stages, and expression of photosynthetic genes increases with maturity. Regulatory genes likely to be involved in C4 photosynthesis are also discussed.

Discussion

The present study demonstrates that challenges associated with de novo transcriptome assembly from short-read sequencing data can be overcome. We establish a workflow for differential expression studies in non-model species using RNA-seq. Using development in Cleome gynandra as an example, we show the approach to be a cost-effective and robust way to combine transcript discovery and transcriptome analysis that can be applied to a wide range of organisms.

Presenting author

Mali Salmon-Divon (mali@ebi.ac.uk)
EMBL-EBI

Author Affiliations

1. EMBL European Bioinformatics Institute, Wellcome Trust Genome Campus, Cambridge 2. Department of Plant Sciences, Downing Street, University of Cambridge, Cambridge, CB2 3EA, UK

A-66. iCount – comprehensive analysis of iCLIP data

*Rot G (1), König J (2), Gorup C (1), Zupan B (1), Ule J (2), Curk T (1, *)*

Compared to other UV-CrossLinking and ImmunoPrecipitation (CLIP) approaches, individual-nucleotide resolution CLIP (iCLIP) that was recently developed (König et al, NMSB 2010) enables a more precise determination of protein-RNA binding sites. Random barcodes, introduced in the DNA adaptor before PCR amplification, are crucial to discriminate between unique cDNAs and PCR duplicates, hence improving the quantification of the protein-RNA interactions. A dedicated computational pipeline was needed to support the processing and analysis of the sequence data gathered in iCLIP experiments.

Materials and Methods

The computational pipeline for the analysis of iCLIP data is implemented in Python. Mapping of reads is performed with Bowtie (Langmead et al., Gen. Biol. 2009). After removing experiment multiplexing and random barcodes, reads are mapped to the genome either by allowing single or multiple hits. Hits from the same sequencing run are filtered together. Reads with detected errors (mismatch, insertion or deletion) are completely removed. Reads with identical random barcodes are collapsed into one read (cDNA). A peak-finding algorithm uses the resulting cDNA counts to identify crosslink clusters.

Results

We present a pipeline that takes raw iCLIP sequence data for input and proceeds in multiple steps to identify the protein-RNA binding sites and quantify the extent of binding. Multiplexed experiments are annotated and binding is quantified based on random barcodes. We show an increased reproducibility of sites with high cDNA count. Sequence 5-mer motifs enriched in proximity of binding sites and within clusters are identified. We show that crosslink clusters represent high-affinity binding sites. Results are available as BED and bedGraph files, and can be explored in a genome browser.

Discussion

The pipeline has been successfully applied for the analysis of multiple proteins. We plan to add functionalities for the downstream analyses of iCLIP data, such as analysis of spatial relations among binding sites of different proteins, changes in binding of a protein under different cellular conditions, generation of RNA-binding maps in proximity of genomic landmarks such as exons and UTRs, and GO term enrichment analysis of bound genes. We are developing an intuitive and easy-to-use interface that will accelerate the discovery process of scientists using the iCLIP experimental method.

Presenting author

Tomaz Curk (tomaz.curk@fri.uni-lj.si)
University of Ljubljana, Faculty of Computer and Information Science

Author Affiliations

1 University of Ljubljana, Faculty of Computer and Information Science, Trzaska cesta 25, SI1000 Ljubljana, Slovenia 2 MRC Laboratory of Molecular Biology, Hills Road, Cambridge CB2 0QH, UK

Acknowledgements

The work was supported by Human Frontiers Science Program grant RGP0024 and the European Research Council grant 206726-CLIP and Slovenian Research Agency (ARRS).

URL

`http://www.ailab.si`

A-68. Exploring 3D pooled next-generation sequencing using SNP-Cub^3

De Schrijver J (1,), De Wilde B (2), Trooskens G (1), Vandesompele J (2), Van Criekinge W (1)*

Next-generation amplicon resequencing allows rapid identification of novel or known SNPs and mutations. However, sample preparation remains a largely manual and sometimes tedious task, especially when large pools of samples need to be analyzed. Multidimensional pooling is a technique where a sample is duplicated in a combination of pools in such a way that every sample is included in a unique combination of pools. This pooling technique allows a drastic reduction of sample pools (from x^y samples to x*y pools, for example 125 samples to 15 pools) but complicates downstream data analysis.

Materials and Methods

The pipeline was developed using MySQL and (Bio)Perl.

Results

A drawback of the pooling approach is that variant frequencies in a single pool are very low. Stochastic effects can make it difficult to classify this as a variant or a sequencing/PCR error. The pipeline is able to overcome this limitation by comparing variant frequencies in several pools wherein a single sample is included. A real sample variant (unique to that sample) in a 3D-pooled setup should appear in 3 pools with the same frequency. SNP-cub^3 is able to assign a p-value, which is the probability that a real variant would be randomly sampled at the observed frequencies.

Discussion

The pipeline allows accurate identification of variants using the pooling strategy. There is also a simulation module available that allows a user to simulate the possibilities of the pipeline and allows optimization of the design of the pool sizes.

Presenting author

Joachim M.C.T. De Schrijver (joachim.deschrijver@ugent.be)
University Ghent

Author Affiliations

1 Laboratory for Bioinformatics and Computational Genomics, Department of Molecular Biotechnology, Ghent University, 9000 Ghent, Belgium 2 Center for Medical Genetics, Ghent University Hospital, 9000 Ghent, Belgium

B-44. An ancient motif signature defines hundreds of vertebrate developmental enhancers

Parker H (1), Piccinelli P (1,), Elgar G (1)*

Gene regulation by cis-regulatory elements plays a crucial role in development and disease. A major aim of the post-genomic era is to be able to read the function of cis-regulatory elements through scrutiny of their DNA sequence. The role of cis-regulatory elements in vertebrate development is demonstrated by the increasing number of examples where mutations lead to genetic diseases. Whilst progress has been made identifying key motifs within vertebrate promoter elements, CNEs have remained recalcitrant to systematic motif-identification approaches, despite some elegant targeted approaches.

Materials and Methods

For the evolutionary analysis we created 2 different test sets of multiple alignments; one jawed vertebrate set and one lamprey set. To find evolutionary conserved Pbx-hox motifs we employed the software Cis-Finder and/or rnabob on our 2 alignment sets and their respective control. A motif match was only considered if it matched all aligned species and occurred at the exact same aligned position. In parallel we also employed a de-novo motif finding strategy implemented in Cis-Finder. It scans a set of DNA sequences for over-represented position frequency matrices, cluster these and rank them.

Results

We searched for instances of the canonical pbx-hox motif - 'TGATNNAT' - conserved across CNE multiple sequence alignments. In a set of 279 alignments of CNEs between human, zebrafish, fugu and lamprey, we identified 79 conserved motifs, representing significant enrichment compared to control sets (2.5-fold zero-order markov, 25-fold alignment shuffling). We complemented our 'bottom-up' search for pbx-hox motifs in CNEs with a 'top-down' de-novo motif search. Strikingly, the top-scoring motifs matches our consensus pbx-hox motif, extending it to the longer 'KR' 10mer motif.

Discussion

We identify, for the first time, a striking enrichment for Pbx-Hox binding site motifs within CNEs. This represents a crucial first step in systematically de-coding vertebrate developmental enhancers and highlights the utility of the sea lamprey as a model system for investigating vertebrate gene regulation. We predict that the set of putative Hox-responsive cis-regulatory elements identified in this study will be a powerful resource for systematically identifying the transcription factor combinations underlying the functional versatility of vertebrate enhancers.

Presenting author

Paul Guiseppe Piccinelli (ppiccin@nimr.mrc.ac.uk)
National Institute of Medical Research

Author Affiliations

(1) National Institute of Medical Research (NIMR), London, UK

Acknowledgements

Medical Research Council (MRC), London, UK

B-45. Diversity of the chromosomal beta-lactamase in 181 clinical Klebsiella oxytoca isolates. Comparison of the relation between beta-lactamase and gyrase-A sequences

Worning P (1,), Boye K (1), Hansen DS (2)*

Klebsiella are among the top five Gram-negative pathogens in hospital-aquired infections. The chromosomal beta-lactamase gene from Klebsiella oxytoca has been shown to exist in six types (Oxy1 to Oxy6). The six types of the Oxy enzyme have different substrate profile and this could lead to differences in the frequencies of the various Oxy-types. We are investigating if the Danish beta-lactamase sequences fall in clearly separated groups. We have sequenced the gyrase-A gene from a subset of the isolates to see if the gyrA and the beta-lac sequences fall in the same groups.

Materials and Methods

The chromosomal beta-lactamase gene from 181 Klebsiella oxytoca isolates from Danish hospitals was sequenced. The sequences were trimmed to initiate at the start codon, the sequences vary in length from 770 to 835 bp. For 25 of the isolates, 348 bp of the gyrase-A gene were sequenced. Clustalx was used to align and make neigbour joining trees with bootstrap values. Comparison of the beta-lac and the gyrA trees made from the 25 isolates were done by visual inspection. Normalized symmetrical distance and a new SPRIT method was used to evaluate the distance between the two trees.

Results

The 181 Danish Klebsiella oxytoca isolates fall in five clearly separated groups (numbers of isolates) Oxy1(44), Oxy2(103), Oxy4(1), Oxy5(10) and Oxy6(23). The bootstrap values for the five groups together with the type sequences are very high. The Oxy6 group is the tightest of the four groups with more than one member, with half the median intra group distance than Oxy1, Oxy2 and Oxy5, which all have the same median intra group distance. The 25 relatively short (348 bp) gyraseA sequences reproduced the grouping with one exception, Oxy1 and Oxy5 made one group.

Discussion

The fact that beta-lactamase and gyrase-A sequences produce the same tree structure supports the evolutionary reliability of the grouping. An extension of the study with more and longer gyrA sequences could be an interesting way to make the conclusions stronger.

Presenting author

Peder Worning (pederworning@hotmail.com)
Bum Bummelum

Author Affiliations

1) Department of Clinical Microbiology, Hvidovre Hospital, Denmark 2) Department of Clinical Microbiology, Hillerød Hospital, Denmark

B-46. Origin and evolution of the organellar release factors

Duarte I (1,), Huynen M (1)*

Only 2 codon-specific RFs are sufficient for de facto organellar translation termination: RF1 and RF2, but at least 3 other different subfamilies have been described, whose exact function remains elusive. Our detailed analysis sought to integrate different sources of information: localization prediction and experimental studies, phylogenetic distribution, organellar genetic code and sequence structural features, to better describe this superfamily and predict unannotated proteins' function.

Materials and Methods

We used comparative genomics (Sequence Homology Searches, Multiple Sequence Alignment and Phylogeny) to study the organellar RF superfamily in a carefully selected group of Eukaryotic organisms. We analyzed and compared primary sequence domains and motifs, as well as their localization signals and data, phylogenetic distribution, evolutionary origin, and organellar genetic code assignment.

Results

This study reports a new plant specific release factor subfamily, which has lost the 3 functional motifs experimentally described as essential for bona-fide release factor activity. Also, it reports a wrongly annotated plant RF2 subfamily, to be chloroplastic and not mitochondrial. The phylogenetic analyses show an alpha-proteobacterial, cyanobacterial and red algal origin for mitochondrial, chloroplastic and plastidial RFs, respectively. The RFs co-evolution with the mitochondrial genetic code found in the green algae lineage, suggests a possible re-invention of the standard genetic code.

Discussion

Our studies confirm the previously published evolutionary origin for the mitochondrial and plastidial RFs, in accordance with the currently undisputed endosymbiotic origin of these organella. Also, the singular absence of functional sequence motifs found in the proposed plant-specific mitochondrial subfamily, seems to indicate that this protein is not functioning as canonical RF. Hence the motif shuffling and loss seen on each subfamily, hints at a significant functional divergence, which serves to prove that much is still to be uncovered about this process!

Presenting author

Isabel Duarte (i.duarte@cmbi.ru.nnl)
CMBI - NCMLS Radboud University

Author Affiliations

CMBI - NCMLS, Radboud University Nijmegen Medical Centre

Acknowledgements

FCT - Fundacao para a Ciencia e Tecnologia, Portugal PDBC - Programa de Doutoramento em Biologia Computacional do Instituto Gulbenkian de Ciencia, Portugal

B-47. Identification and characterisation of novel "atypical" odorant binding proteins in the mosquito genomes

Manoharan M (1,2,3,), Ng Fuk Chong M (1,2), Vaitinadapoulle A (1,2), Frumence E (1,2), Sowdhamini R (3), Offmann B (1,2)*

The spread of infectious diseases among humans is mediated primarily by the mosquitoes. Their ability to recognize the human host relies on olfactory mechanism and odorant binding proteins (OBPs) are important components of this mechanism. OBPs are small extracellular proteins which help in the solubilisation and transport of odorants from the external environment to their membrane targets triggering signal synapsis. This study revolves around the identification and analysis of these proteins in three available mosquito genomes.

Materials and Methods

The predicted peptide sequences of the three genomes were downloaded from Vectorbase. The putative odorant binding proteins were identified using cascading Psiblast with an E-value of 3e-10 and an alignment length cutoff of 75% with respect to the query sequence. Multiple sequence alignments were constructed using ClustalW based on sequence to profile alignment using a structure based alignment of available structural information of the family from the PDB as the profile. The phylogenetic trees were inferred using the Neighbor-Joining method with 1000 bootstrap replicates in MEGA 4.0.

Results

We report, for the first time, the presence of 26 atypical OBPs in Culex quinquifasciatus and 31 additional members in the Aedes aegypti genome. A cross genome comparison helped us to identify within the atypical OBPs 4 subtypes, which we named Matyp1-4, featured by unique cysteine conservation patterns that were not observed earlier. Interestingly the MAtyp2 members of this cluster (with the exception of Anopheles members) were found to lack C2, C5, C7, C8, C9, C11 making a total of only 6 conserved cysteines while the other types have 12 conserved cysteines.

Discussion

Atypical OBPs have a C-term extension when compared to classic OBPs. Their role and distribution are still to be uncovered. We here report new atypical members in Culinidaes with new structural features. The concerted absence of C2 and C5, characteristic of Minus C OBPs only reported in D. melanogaster so far, has motivated us to name this new type of atypical OBPs as "Minus C like atypical odorant binding proteins". The classification and the cysteine conservation pattern observed makes the atypical members an important class of odorant binding proteins for further functional analysis.

Presenting author

Malini Manoharan (malini.manoharan@univ-reunion.fr)
INSERM, DSIMB UMR-S665, University of La Reunion

Author Affiliations

1 - DSIMB, University of La Reunion, La Reunion, France 2 - DSIMB, INSERM, UMR-S 665, La Reunion, France 3 - National Center for Biological Sciences, GKVK campus, Bangalore, India

Acknowledgements

MM is funded by a PhD scholarship from Conseil Regional de La Reunion. This research is funded by a research grant to BO from Conseil Regional de La Reunion. RS is grateful to University of La Reunion for invited professorship.

URL

`http://bioinformatics.univ-reunion.fr/`

C-41. Different structures/same interaction partners: what can we learn from promiscuous protein-protein interactions?

Martin J (1,)*

The fact that proteins with different 3D structures can bind similar partners suggests that some convergently evolved binding interfaces are re-used in different complexes. Such interfaces, if any, could constitute the tip of the iceberg of an hypothetical limited repertoire of protein-protein binding interfaces common to all protein-protein complexes.

Materials and Methods

In order to characterize these convergently evolved interfaces, we analyzed a set of protein complexes composed of non-homologous domains interacting with homologous partners at equivalent binding sites. We focused on quantifying the extent of physico-chemical similarity of interfaces. We assessed the significance of the similarity by bootstrapping the atomic properties at the interfaces.

Results

We found that the similarity of binding sites is very significant between homologous proteins, as expected, but generally insignificant between the non-homologous proteins that bind to homologous partners. We could only identify a limited number of cases of structural mimicry at the interface, suggesting that this property is less generic than previously thought.

Discussion

Our results support the hypothesis that different proteins can interact with similar partners using alternate strategies, but do not support convergent evolution.

Presenting author

Juliette Martin (juliette.martin@ibcp.fr)
CNRS Institut de Biologie et Chimie des Protéines UMR 5086

Author Affiliations

1: Institut de Biologie et Chimie des Protéines UMR 5086 CNRS Université de Lyon

C-42. 3DM: A system to automatically build structure-based super-family systems

Joosten HJ (1), Kuipers RKP (1,2), Vd Bergh T (1), Schaap PJ (1), Vriend G (2,)*

The 3DM system is widely applicable. The reason for presenting it at the ECCB is that the system answers surprisingly more molecular questions than one would expect from its simplicity.

Materials and Methods

The system revolves around a multi-structure alignment that guides the multi-sequence alignment (MSA). This MSA is thus of very great depth and high quality and allows us to draw conclusions that cannot be drawn using more 'classically derived' MSAs.

Results

The MSA is annotated with automatically extracted mutation data from literature, contact data from available structure files, and computationally derived data, such as conservation and correlated mutations. All data are connected via a unified numering scheme that is based on the 3D-MSA and applied to all data making it easy find hidden correlations between the data and to plot different data-types on the MSA and in the structures.

Discussion

The connectivity of the data within a 3DM system makes that these systems can succesfully be applied in a large series of research projects related to enzyme activity or the understanding and engineering of specificity, protein stability engineering, DNA-diagnostics, drug design, etcetera.

Presenting author

Gert Vriend (vriend@cmbi.ru.nl)
CMBI RUNMC

Author Affiliations

1) Bio-Prodict; Dreijenplein 10; 6703 HB Wageningen; The Netherlands 2) CMBI; Geert Grooteplein 26-28; 6525 GA; Nijmegen; The Netherlands

Acknowledgements

The authors thank NBIC for financial support.

C-43. Quantifying the relationship between RNA sequence and three-dimensional structure conservation for homology detection

Capriotti E (1,2,), Marti-Renom MA (3)*

In the last decade, the number of available RNA structures has rapidly grown reflecting the increased interest on RNA biology. Similarly to the studies carried out in the late eighties for proteins, which gave the fundamental grounds for developing comparative protein structure prediction methods, we are now able to quantify the relationship between sequence and structure conservation in RNA.

Materials and Methods

Here we introduce an all-against-all sequence- and three-dimensional (3D) structure-based comparison of a representative set of 451 RNA structures using the SARA algorithm. Selecting a set of 589 high similarity alignments from 114 RNA chains, we have quantitatively confirm that: (i) there is a measurable relationship between sequence and structure conservation, (ii) evolution tends to conserve more RNA structure than sequence, and (iii) there is a twilight zone for RNA homology detection.

Results

We found that similar to proteins, the structure identity decreased with the decrease of the sequence identity and in agreement with previous work and we observed a higher mean value of percentage of structure identity with respect to the average sequence identity. Using the Infernal program, an e-value threshold of ~5e-4 has been found to be the lower limit of the "twilight zone" for sequence alignment longer than 100 nucleotides.

Discussion

The computational analysis here presented quantitatively describes the relationship between sequence and structure for RNA molecules and defines a twilight zone region for detecting RNA homology. Our work could represent the theoretical basis and limitations for future developments in comparative RNA 3D structure prediction.

Presenting author

Emidio Caprotti (emidio@stanford.edu)
Department of Bioengineering, Stanford University

Author Affiliations

1 Department of Bioengineering, Stanford University, Stanford (CA), United States of America. 2 Department of Mathematics and Computer Sciences, University of Balearic Islands, Palma de Mallorca, Spain. 3 Structural Genomics Unit. Bioinformatics and Genomics Department. Centro de Investigación Príncipe Felipe. Valencia. Spain

Acknowledgements

EC acknowledges support from the Marie Curie International Outgoing Fellowship program (PIOF-GA-2009-237225). MAM-R acknowledges support from the Marie Curie International Reintegration program (FP6-039722), the Generalitat Valenciana (GV/2007/065), and the Spanish Ministerio de Ciencia e Innovación (BIO2007/66670).

C-44. IPknot: fast and accurate prediction of RNA secondary structures with pseudoknots using integer programming.

Sato K (1,), Kato Y (2), Akutsu T (2), Asai K (1,3)*

Pseudoknots, substructures observed in RNA secondary structures, play a role in assisting the overall 3D folding in many cases, and thus prediction of RNA secondary structures including pseudoknots is expected to provide a clue to determine the 3D structures of RNA molecules.

Materials and Methods

IPknot executes the following two steps when an RNA sequence is given: (1) approximate a posterior probability distribution over a space of pseudoknotted secondary structures by its factorization, each of which is defined over a space of pseudoknot-free secondary structures; (2) solve the integer programming (IP) problem to predict the optimal secondary structure in terms of maximizing expected accuracy. We introduce a threshold cut technique to solve the IP problem efficiently, which is shown to be reasonable from the viewpoint of maximizing expected accuracy.

Results

We selected a dataset from the RNA STRAND database on condition that a test sequence includes at least one pseudoknot, the length is at most 500 bp and each sequence has sequence similarity less than 85 %. IPknot was compared in prediction performance with ILM, pknotsRG, FlexStem, RNAfold and CentroidFold, showing that IPknot outperforms the competitive methods on the dataset. Furthermore, IPknot runs faster than the other prediction methods except ILM.

Discussion

Although IPknot yields good prediction on relatively long sequences, the method shows a falling tendency in accuracy when dealing with short sequences, which is due to the fail in detecting pseudoknots. To remove this drawback, tuning parameters of IPknot appropriately is necessary, which is our ongoing work.

Presenting author

Kengo Sato (satoken@k.u-tokyo.ac.jp)
University of Tokyo

Author Affiliations

1: University of Tokyo, Japan, 2: Kyoto University, Japan, 3: National Institute of Advanced Industrial Science and Technology, Japan.

Acknowledgements

This work was supported by Grant-in-Aid for Young Scientists (B) (KAKENHI) from Ministry of Education, Culture, Sports, Science and Technology (MEXT), Japan [#22700305 to K.S., #22700313 to Y.K.].

C-45. The Protein Model Portal, a resource of the Nature PSI Structural Biology Knowledgebase

Bordoli L (1,2), Haas J (1,2), Benkert P (1,2), Mostaguir K (1,2), Kiefer F (1,2), Arnold K (1,2), Schwede T (1,2,)*

The Protein Model Portal (PMP) has been developed to foster effective use of molecular models in biomedical research by providing convenient and comprehensive access to structural information for a specific protein. Both experimental structures and theoretical models for a given protein can be searched and simultaneously, and analyzed for structural variation. PMP is a resource of the Nature PSI (Protein Structure Initiative) Structural Biology Knowledgebase (sbkb.org).

Materials and Methods

PMP is a portal, federating experimental protein structure information and structure models from different resources: the PDB, the PSI (Protein Structure Initiative) centers - CSMP, JCSG, MCSG, NESG, NMHRCM, NYSGXRC, JCMM – and the comparative modeling resources ModBase and SWISS-MODEL Repository. PMP also provides access to interactive services for model building (ModWeb, M4T, SWISS-MODEL) and model quality estimation (ModFOLD, QMEAN).

Results

The current release of the PMP (August 2010) consists of 13.5 million comparative protein models for 3.5 million distinct UniProt sequences. About one quarter of the residues of all the UniProt sequences are covered by at least one model accessible from the Protein Model Portal. Model quality estimation tools allow the user evaluating the accuracy of the generated models to indicate their usefulness for specific applications in biomedical research.

Discussion

Through PMP for the first time it is now possible to query all participating federated structure resources simultaneously and compare the available structural models in a single interface. Tools for structure comparison immediately highlight structurally conserved segments and regions of structural variation. Theoretical models generated using various algorithmic approaches with different strengths and weaknesses can be assessed using the same model quality estimation tools to select the most appropriate one.

Presenting author

Torsten Schwede (torsten.schwede@unibas.ch)
Biozentrum, University of Basel and Swiss Institute of Bioinformatics, Basel, Switzerland

Author Affiliations

(1) Biozentrum, University of Basel, Basel, Switzerland, (2) SIB, Swiss Institute of Bioinformatics, Basel, Switzerland

Acknowledgements

PMP is supported by the National Institutes of Health NIH as a sub-grant with Rutgers University, under Prime Agreement Award Number: 3U54GM074958-04S2

URL

http://schwedelab.org

C-46. PB-PENTAdb : a web resource for the analysis and prediction of local backbone structure and flexibility using pentapeptide protein blocks

Offmann B (1,2,3,), Tyagi M (4), Joseph A (5,6) Drula M (1,2), Grondin M (1,2), Frumence E (1,2), Vaitinadapoulle A (1,2), Cadet F (1,2,3), Srinivasan N (7), de Brevern A (5,6)*

Little is known about the structural diversity of pentapeptides in protein structures. Since the inspiring Kabsch and Sander paper in 1983 and Argos in 1987, there was no equivalent update of their work. Taking advantage of protein blocks, a 16-state description of the backbone of proteins, we have (re)investigated the sequence to structure relationship at the pentapeptide level but on a much larger scale than for the 1987 article and explored potential applications for the analysis and prediction of local backbone structure and flexibility.

Materials and Methods

A subset of SCOP 1.75 domains referred as ASTRAL100 with sequences culled at 100% sequence identity was used as initial data. Other datasets with higher quality and varying degree of non-redundancy were also used. Structural domains were converted into 1D representations in terms of protein blocks (PBs) using de Brevern's algorithm. From these, pentapeptide sequences and their local structural status were extracted and stored in a database. We developed a flexibility index inspired from Simpson's diversity index and a knowledge-based prediction tool to predict local structure of proteins.

Results

A database containing 5.73 million pentapeptides and their local structures was derived from 33479 SCOP domains from ASTRAL100. The total sequence space covered is 1.28 million out of the 3.2 million possible pentapeptides. About 32% of these occurred only once in the database and 40% occurred at least 4 times. Strong and weak sequence-structure relationships i.e "rigid" and "flexible" pentapeptides could be identified very easily from our database. Amino acid usage and biological context of these structural "hotspot" are discussed.

Discussion

We are providing for the first time, an online tool for the assessment of the structural diversity of pentapeptides in terms of PBs, a 16-state description of the backbone. The statement that "identical pentapeptides can have completely different conformations…" is still true today and our work demonstrates it very clearly ; backbone plasticity is a rule in the pentapeptide sequence space. Nevertheless, there are hundreds of pentapeptides that do not change conformation though found in different biological contexts. Theses "rigid" pentas can provide clues on sequence-structure relationships.

Presenting author

Bernard G Offmann (bernard.offmann@univ-reunion.fr)
University of La Reunion

Author Affiliations

1 - DSIMB, University of La Reunion, La Reunion, France 2 - DSIMB, INSERM, UMR-S 665, La Reunion, France 3 - Peaccel, Cambridge, MA, USA 4 - NCBI, NIH, Bethesda, MD, USA 5 - DSIMB, INSERM, Paris, France 6 - DSIMB, INTS, Paris, France 7 - Molecular Biophysics Unit, Indian Institute of Science, Bangalore, India

Acknowledgements

This research is in part funded by Conseil Regional de La Reunion

URL

`http://bioinformatics.univ-reunion.fr/`

C-47. RactIP: fast and accurate prediction of RNA-RNA interaction using integer programming

Kato Y (1,), Sato K (2), Hamada M (3,4), Watanabe Y (5), Asai K (2,4), Akutsu T (1)*

Considerable attention has been focused on predicting RNA-RNA interaction since it is a key to identifying possible targets of non-coding small RNAs that regulate gene expression post-transcriptionally. A number of computational studies have so far been devoted to predicting joint secondary structures or binding sites under a specific class of interactions. In general, there is a trade-off between range of interaction type and efficiency of a prediction algorithm, and thus efficient computational methods for predicting comprehensive type of interaction are still awaited.

Materials and Methods

RactIP executes the following two steps when two RNA sequences are given: (1) approximate a posterior probability distribution over a space of joint secondary structures by its factorization, using internal and external base-pairing probabilities; (2) solve the integer programming (IP) problem to predict the optimal joint structure in terms of maximizing expected accuracy. We introduce a threshold cut technique to solve the IP problem efficiently, which is shown to be reasonable from the viewpoint of maximizing expected accuracy.

Results

We first conducted experiments in joint secondary structure prediction on five sRNA-target pairs. Results show that RactIP outperforms inteRNA in accuracy and is comparable to inRNAs (the exact model). Moreover, RactIP runs overwhelmingly faster than the competitive methods. In the second experiment in binding site prediction, accuracy of RactIP is as good as that of inRNAs (the heuristic model) and better than that of IntaRNA on 23 sRNA-target pairs.

Discussion

RactIP is expected to improve prediction performance in unknown target search in long genomes by predicting respective intramolecular structures as well as intermolecular binding sites in practical time. For this purpose, we should improve RactIP so that it can discriminate between targets and non-targets, which is left as our future work.

Presenting author

Yuki Kato (ykato@kuicr.kyoto-u.ac.jp)
Kyoto University

Author Affiliations

1: Kyoto University, Japan, 2: University of Tokyo, Japan, 3: Mizuho Information & Research Institute, Inc, Japan, 4: National Institute of Advanced Industrial Science and Technology, Japan, 5: Doshisha University, Japan.

Acknowledgements

This work was supported by Grant-in-Aid for Young Scientists (B) (KAKENHI) from Ministry of Education, Culture, Sports, Science and Technology (MEXT), Japan [#22700313 to Y.K., #22700305 to K.S.].

D-27. Functional Inference for Experimental Proteomics: The PANDORA (Annotations graph) and ProtoNet (Family Tree) Resources

Rappoport N (1,), Linial M (2,3)*

The study of proteins and their properties is essential for medicine, agriculture and biological research. An intimate knowledge on proteins is usually based on combining technologies. The number of protein sequences that are currently known already reaches 10 millions. With the recent advances in genomic and proteomics, this number keeps growing. Consequently, it becomes increasingly hard to gain a global view on protein functions without the use of appropriate bioinformatics tools. Automatic functional inference and visualization tools indeed become essential component for this task.

Materials and Methods

We developed unsupervised large-scale databases and tools for clustering proteins to their homologous groups. ProtoNet is based on 'all against all' BLAST results that are used to construct a hierarchy tree of proteins. ProtoNet allows detecting overlooked evolutionary connections. PANDORA is an integrative platform that covers proteins structure, domains, taxonomy, ontology and properties of the data as provided by the users. PANDORA visualizes protein sets as directed acyclic graphs (DAG) built from the shared annotations of the proteins in the analyzed set.

Results

ProtoNet provides a set of tools for comprehensive analysis of all protein sequences in the current UniProt resource. It allows a dynamic view of protein clusters at different levels of granularity and according to varying rules. When studying a certain protein family, analyzing clusters according to the different merging rules captures evolutionary information. The protein set that are analyzed by PANDORA is now selected by multiple modes. Most recent addition is the support of MS-base proteomics. PANDORA supports proteolytic peptides and assigns them to the identified proteins.

Discussion

While it is expected that the huge body of genomes will accelerate the success in inferring proteins structure and function, over 50% of eukaryotes predicted ORFs are still lacking assigned functions. Our methods and resources are important components in the task of automatic functional prediction. Evaluating the power of prediction methods at a genomic scale using our tools is unique as the analysis performed on protein collections rather than on individual selected proteins. Such collections are a natural outcome of experimental and high throughput technologies.

Presenting author

Nadav Rappoport (nadavrap@cs.huji.ac.il)
School of Computer Science and Engineering, The Hebrew University of Jerusalem, Israel

Author Affiliations

School of Computer Science and Engineering and Department of Biological Chemistry, Institute of Life Sciences, The Hebrew University of Jerusalem, The Sudarsky Center for Computational biology, Jerusalem, 91904 Israel.

Acknowledgements

This work was supported by EU Framework VII Prospects consortium and a grant from ISF 592/07.

D-28. Computational approaches to study glycosaminglycans recognition by IL-8 for the design of biomaterials for tissue regeneration

Samsonov SA (1,) , Pisabarro MT (1)*

Interactions of proteins with glycosaminglycans (GAGs) determine many mechanical and regulatory cell properties. However, computational studies of GAGs-protein interactions are challenging because of GAGs high conformationally flexibility and symmetry, and a lack of well-established theoretical approaches to treat GAGs and their water-mediated interactions. We analyze interactions of the chemokine IL-8 and different GAGs that underly the regulation of tissue regeneration to support the design of biomaterials that promote bio-specific cell behaviour.

Materials and Methods

We apply a docking approach to propose poses for 14 tetrameric derivatives of chondroitin sulfate and hyaluronic acid binding to monomeric IL-8. Molecular dynamics (MD) with GLYCAM06 force field implemented in AMBER10 and binding free energy calculations using MM-PBSA methodology are used for refinement and further dynamical and energetical characterization of the binding poses. For our MD calculations we create sulfated GAGs units libraries compatible with GLYCAM06.

Results

For all docked GAGs we find one common highly scored binding pose, which also demonstrates stable behaviour in the MD simulations. We decompose binding energy per IL-8 residue to determine the impact of individual residues to binding. We also observe that binding energy and mobility of sugars highly depend on the degree of GAGs sulfation. Increase of the GAGs length does not linearly improve energy of binding with a monomeric IL-8. Association of IL-8 monomeric units into a dimer is significantly affected by GAGs binding. This effect increases together with the length of the bound GAG.

Discussion

Our results show that electrostatics play a decisive role in GAGs-IL-8 interactions, while position-specific GAGs sulfation impact is rather challenging to be detected by MM-PBSA. IL-8 GAGs binding site essentially determines enthalpy, whereas increase of the length of GAGs contributes to entropy of GAGs binding. GAGs binding influences dimerization of IL-8, which is of extreme importance for biological function of IL-8. The obtained data deepens our understanding of GAGs recognition and are being used for rational engineering of biomaterials in studies of skin and bone tissues regeneration.

Presenting author

Sergey A. Samsonov (sergeys@biotec.tu-dresden.de)
BIOTEC TU Dresdem

Author Affiliations

1. Structural Bioinformatics Group, BIOTEC TU Dresden, Tatzberg 47/49, 01307 Dresden, Germany.

Acknowledgements

This work was funded (in part) by the German Research Council SFB-TRR 67.

D-29. Expression profile based substrate specificity prediction in complete Saccharomyces cerevisiae methyltransferome

Wlodarski T (1,2,),Rowicka M (3), Ginalski K (1)*

Methylation is one of the most important modification of proteins, nucleic acids and small molecules. Its complete description is crucial to understand many cellular processes, such as signal transduction or transcriptional control. By characterizing all known and putative methyltransferases (MTases) together with finding so far unknown members, we have described the complete MTase world in S. cerevisiae. This genome wide study fits into current trends in bioinformatics and could be used as a starting point in deeper analysis of methylation and its relation to other crucial cellular processes

Materials and Methods

The yeast methyltransferome was completed with novel MTases detected with Meta-BASIC method for distant homology detection. All yeast MTases were classified structurally, based on fold assignment of their catalytic domains using 3D-Jury server and SCOP classification. For putative and novel MTases, we have developed a new approach for substrate specificity prediction. It is based on correlation of MTases expression profiles in Yeast Metabolic Cycle with substrate specificity of known MTases, combined with hierarchical clustering and decision trees

Results

Identified methyltransferome includes 52 known MTases, 33 putative (previously annotated, yet without biochemically confirmed activity) and one previously unknown. 56 MTases retain Rossmann fold - the most common structural fold for MTases, while the remaining can be classified into 8 other folds. Moreover, we have observed high similarity between expression profiles of all RNA MTases, which was used together with calculated isoelectric point for substrate specificity prediction. Using this new approach we have predicted substrate specificity for 24 out of 33 putative and for one novel MTase

Discussion

We have completed and characterized proteins involved in one of the most important biochemical process in all living cells – methylation. Additionally, for the first time we have observed high similarity between expression profiles of all RNA MTases and have developed new methodology for their substrate specificity prediction.

Presenting author

Tomek Wlodarski (didymos@icm.edu.pl)
ICM, University of Warsaw

Author Affiliations

1.Univ Warsaw, Interdisciplinary Ctr Math & Computat Modelling, Warsaw, Poland 2.Univ Warsaw, Inter-Faculty Interdisciplinary Doctoral Studies in Natural Sciences and Mathematics, Warsaw, Poland 3. University of Texas Medical Branch, Department of Biochemistry & Molecular Biology,Galveston, USA

D-30. Discovery and annotation strategies of nonribosomal peptide synthetases from bacterial genomes

Saci Z (1,2), Pupin M (1,), Deravel J (2), Krier F (2), Caboche S (1,2), Jacques P (2), Leclère V (2)*

Nonribosomal peptide synthetases (NRPSs) are huge multi-enzymatic complexes synthesizing peptides, but not through the classical process of transcription and then translation. The synthetases are organised in modules, each one integrating an amino acid in the final peptide. The modules are divided in domains providing specialized activities. So, those enzymes are as diverse as their products and Norine, the database dedicated to nonribosomal peptides, listed more than 1000 peptides. We present our strategy to annotate them accurately and its application to the Erwiniae genomes exploration.

Materials and Methods

We use three ways to identify proteins likely to be synthetases: keyword search among available annotations, selection of huge proteins and BLASTP search. Selected proteins are then studied with bioinformatics tools dedicated to NRPSs (NRPS-PKS, Ansari et al., NAR 04 ; NRPSpredictor, Raush et al., NAR 05 ; PKS/NRPS Analysis Web-site, Bachmann and Ravel, Methods Enzymol. 09). Those tools and our expertise allowed us to reconstruct the modules composing the synthetases, predict the amino acids they potentially integrate and the activity of the produced peptide with the help of Norine.

Results

A systematic search for NRPSs genes on available genomic sequences was done on several bacterial genus and species. This work emphasizes the bad quality of automatic annotations. Some annotations are imprecise such as "putative NRPS". In this case, they are correct even if they are not standardized. But, when they are more precise, they are frequently incorrect, for example the number of modules is erroneous or the product name is wrong. Other synthetases are annotated as putative or hypothetical proteins. Those observations are true for all the genomes explored.

Discussion

To conclude, the growth of genome sequences available in the databanks is going with a decrease of the annotation quality. This observation is particularly true for genes coding for modular enzymes such as NRPSs. A more meticulous analysis is necessary for their annotation and a strategy is proposed by this work. We annotate several dozens of synthetases genes in various microbial genomes including Erwiniae.

Presenting author

Maude Pupin (maude.pupin@lifl.fr)
Sequoia, INRIA Lille - Nord Europe and LIFL (UMR 8022)

Author Affiliations

1. SEQUOIA, LIFL (UMR Lille 1/CNRS 8022), INRIA Lille - Nord Europe 2. ProBioGEM (UPRES EA 1026), Univ Lille Nord de France

Acknowledgements

This work was supported by the PPF bioinformatique program of Lille 1 University and INRIA. The ProBioGEM lab is supported by the Region Nord-Pas-de-Calais, the Ministere de l'Enseignement Superieur et de la Recherche, the French ANR Agency, and European Funds for Regional Development.

D-31. Predicting N-glycosylation sites in human proteins

Gupta R (1), Frederiksen J (1,), Jung E (2), Brunak S (1)*

N-glycosylation is an important post-translational modification that is important in protein folding, cell-cell interaction and cell-recognition. Despite popular belief, the acceptor sites for N-linked glycosylation are not completely characterised. The consensus sequence, N-X-[ST] (X!=P) is known to be a prerequisite for N-glycosylation. However, only two-thirds of these sequons are typically modified. The aim of this project is to predict, using protein sequence and structural features, which sequons would be glycosylated. We also study impact of N-glycosylation on disease mutations.

Materials and Methods

Artificial Neural Networks were trained across a sequence window around the acceptor asparagine to distinguish occupied and unoccupied sequons. The dataset used to train the N-glycosylation predictor, NetNGlyc, consisted of 373 human proteins that had confirmed N-glycosylated sites. The spread of N-glycosylation sites in the human proteome was investigated to determine whether certain categories of proteins or positions along the protein length were more prone to N-glycosylation.

Results

Cross-validated performance showed that the networks could identify 86% of the glycosylated and 61% of the non-glycosylated sequons, with an overall accuracy of 76%. N-glycosylation sites occurred more in protein functional groups such as 'Protein binding'. The sites were revealed to cluster on the N-terminal to the center (10-50% of the protein length) of the protein chain.

Discussion

The interest in N-glycosylation is increasing in both Eukaryotes and Prokaryotes. The impact of N-glycosylation on SNPs and mutations and correlations with disease is a useful investigation to aid in our understanding of SNP linked diseases. N-glycosylated proteins are showed to mainly regulate processes within the cell and belong to the functional category of 'Transport and Binding'. NetNGlyc is useful in reducing the number of consensus sequons while looking for potentially glycosylated sites, it is also shown to predict 80% of the ca.6000 glycosylated sites found in a recent mouse study.

Presenting author

Juliet W. Frederiksen (julietwf@cbs.dtu.dk)
Center for Biological Sequence Analysis, Technical University of Denmark

Author Affiliations

1) Center for Biological Sequence Analysis, Technical University of Denmark. 2) Swiss Institute of Bioinformatics, Lausanne, Switzerland.

D-32. AIGO: Toward a unified framework for the Analysis and the Inter-comparison of GO functional annotations

Hindle M (1,2,), Powers S (3), Habash D (3), Saqi M (2), Defoin-Platel M (2)*

Given the advent of high-throughput sequencing technologies and the resulting data explosion, an urgent requirement to provide electronically inferred functional annotations for sequences of unknown function has given rise to a large number of cross-species annotation inference pipelines. Evaluation of the quality of the resultant functional annotation is challenging given there is often no existing gold-standard to evaluate precision and recall against.

Materials and Methods

We propose of set of nine metrics and implement them in a Python open source library called AIGO for the Analysis and the Inter-comparison of GO functional annotations. The usefulness of the library is demonstrated by analyzing and comparing the publicly available annotations of two Affymetrix microarrays.

Results

Results indicate surprising divergence between the different functional annotations tested, both in terms of the quantity and the quality of the annotations.

Discussion

The divergence of annotations can be caused by variations in a priori information, methodology, and parameters of the prediction methods. These influences vary between annotation pipelines, over time as methods and data sources change, and within the pipeline parameter space. In this work we present a general and independent framework that moves toward a standardized tool set for evaluating the relative effects of these changes on the annotation set composition.

Presenting author

Michael Defoin-Platel (michael.defoin-platel@bbsrc.ac.uk)
Rothamsted Research

Author Affiliations

1 Multidisciplinary Centre for Integrative Biology, The University of Nottingham, UK 2 Department of Biomathematics and Bioinformatics, Rothamsted Research, Harpenden, UK 3Department of Plant Science, Rothamsted Research, Harpenden, UK

E-59. Towards real-time control of gene expression: Controlling the HOG signalling cascade

Uhlendorf J (1,2), Bottani S (2), Fages F (1), Hersen P (2), Batt G (1,)*

To cope with the complexity of understanding biomolecular systems, quantitative models are increasingly needed. Obtaining a quantitative description of biological processes necessitates the capability to observe the response of biological systems subjected to large numbers of different perturbations. However, because of the difficulty to perturb precisely a biological system in a dynamical manner, most studies simply use step-response, static perturbations. In contrast, time-varying perturbations have the capacity to provide much richer information on the system dynamics.

Materials and Methods

To improve our capacity to express in vivo a protein of interest in a chosen time-dependent way, and thus perturb biological systems, we propose to develop a platform for the real-time control of gene expression. The platform integrates a microfluidic device that allows modulating gene expression by changing the extracellular environment, a fluorescent microscope that allows monitoring gene expression, and computational approaches that in real time compute the inputs to apply to the system to obtain the desired outputs.

Results

In this work, we present preliminary results on the control of the HOG signalling cascade in yeast using our platform equipped with a simple PID controller, and on the development more elaborate, model based control approaches. To the best of our knowledge, this is the first application of control theory towards the actual control of gene expression at the single-cell level.

Presenting author

Gregory Batt (gregory.batt@inria.fr)
INRIA

Author Affiliations

1: Contraintes research group, Institut National de Recherche en Informatique et en Automatique, INRIA Paris-Rocquencourt, France 2: Laboratoire Matière et Systèmes Complexes, Université Paris Diderot and Centre National de la Recherche Scientifique, UMR 7057, Paris, France

E-60. Sequencing the transcriptome of a deep-sea hydrothermal vent mussel: new possibilities for the discovery of immune genes in an unconventional model organism

Bettencourt R (1,), Stefanni S (1), Pinheiro M (2), Egas C (3), Serrão Santos R (1)*

Bathymodiolus azoricus is a deep-sea hydrothermal vent mussel found in large faunal communities living in chemosynthetic environments at the sea floor near the Mid-Atlantic Ridge. In an attempt to understand physiological responses of vent mussels to hydrothermal vent conditions, and to reveal genes potentially involved in innate immunity we carried out a high-throughput transcriptome sequence analysis of freshly collected B. azoricus gills tissues as the primary source of immune transcripts given its strategic role in filtering the surrounding waterborne potentially infectious microorganisms.

Materials and Methods

Gills were processed for total RNA and mRNA purifications with the RiboPure™ kit and Poly(A)Purist™, respectively. The normalized cDNA (SMART methodology) was sequenced with the 454 GS Titanium platform (Roche). Quality reads were subjected to the MIRA assembler. The translation frame of the contigs was determined through searches against the NCBI nr database using BLASTx with an E-value of 10-6 and E-value of 10-2. The software ESTScan detected further potential transcripts. The collection of sequences was processed by InterProScan and Gene Ontology. Results were compiled into a SQL database.

Results

A normalized cDNA library from gills tissue was sequenced in a full 454 GS-FLX run. Assembly of the high quality reads resulted in 75,407 contigs of which 3,071 were singletons. A total of 39,425 transcripts were conceptually translated into amino-sequences of which 22,023 matched known proteins in the NCBI nr protein database, 15,839 revealed conserved protein domains through InterPro functional classification and 9,584 were assigned with Gene Ontology terms. Queries conducted within the database enabled the identification of genes putatively involved in immune and inflammatory reactions.

Discussion

We have established the first tissue transcriptional analysis of a deep-sea hydrothermal vent animal and generated a substantial EST data set from which a comprehensive collection of genes coding for putative proteins was organized in the dedicated database, DeepSeaVent, the first deep-sea vent animal transcriptome database based on the pyrosequencing technology. This provides the most comprehensive sequence resource for identifying novel genes currently available for a deep-sea vent organism, particularly, genes putatively involved in immune and inflammatory reactions in vent mussels.

Presenting author

Raul Bettencourt (raul@uac.pt)
IMAR-Center. University of the Azores

Author Affiliations

(1) University of the Azores, 9901-861 Horta, Portugal. (2)Bioinformatics Unit, Biocant, 3060-197 Cantahede, Portugal. (3) Advanced Services, DNA sequencing Unit, Biocant, 3060-197 Cantahede, Portugal.

Acknowledgements

We acknowledge the Portuguese Foundation for Science and Technology, FCT-Lisbon and the Regional Azorean Directorate for Science and Technology, DRCT-Azores for pluri-annual and programmatic PIDDAC and FEDER funding to IMAR/DOP Research Unit #531 and the Associated Laboratory #9 (ISR-Lisboa); The network of Excellence MarBEF (Marine Biodiversity and Ecosystem Functioning-contract Nº GOCE-CT-2003-505446); the Luso-American Foundation FLAD (Project L-V-173/2006); the Biotechnology and Biomedicine Institute of the Azores (IBBA), project M.2.1.2/I/029/2008-BIODEEPSEA and the FCT: project PTDC/MAR/65991/2006 IMUNOVENT under the coordination of RB.

E-61. Towards a computational tool to uncover genes involved in signaling crosstalk in Arabidopsis thaliana

Omranian N (1,2,), Arvidsson S (1,2), Riaño-Pachón DM (1,2,3), Mueller-Roeber B (1,2)*

There is a wealth of information waiting to be discovered from microarray experiments deposited in public databases. Although these experiments have been carried out with various experimental setups, most analyses have so far focused on single conditions. We aim at characterizing gene expression across different experiments to discover previously unknown links between signaling networks for a better understanding of cellular dynamics. A tool for the identification of cross-talking genes will be useful for experimentalists, guiding them in hypothesis generation and experimental design.

Materials and Methods

Microarray expression data and meta-information were gathered from the Gene Expression Omnibus (GEO). The meta-data of the experiments were annotated against the developmental stage, structure and environment ontologies for Arabidopsis thaliana using the full-text search in MySQL and the NCBO BioPortal ontology annotation web service (http://bioportal.bioontology.org/annotator). Biclustering was performed using the SAMBA algorithm and functional mapping of genes was performed on each bicluster using BioMaps (http://www.virtualplant.org).

Results

We gathered all microarray experiments for Arabidopsis thaliana from GEO, and were able to annotate almost all (97%) of them with at least one ontology term. The algorithm iteratively uses biclustering and functional mapping of genes to identify the most important genes involved in several signaling pathways. To validate our signaling crosstalk discovery method, we carried out the analysis for nitrate and hormone treatments and compared the results with those presented in a previous study (Nero et al., BMC Systems Biology 2009, 3:59), and we could reproduce 85% of their results.

Discussion

Having optimized the parametrization, we are now working on extending our approach to cover all ontology terms that could be annotated to the microarray metadata, and designing a user interface to provide experimental scientists with a useful service in discovering genes involved in signaling crosstalk for selected treatments. The tool will be helpful in experimental design and hypothesis generation. We are also exploring the use of association rule mining as an alternative to biclustering, which would allow the explicit formulation of logical rules for assigning genes to clusters.

Presenting author

Nooshin Omranian (omranian@uni-potsdam.de)
University of Potsdam

Author Affiliations

1- University of Potsdam, Institute of Biochemistry and Biology, Karl-Liebknecht-Str. 24-25, 14476 Potsdam, Germany 2- Max Planck Institute of Molecular Plant Physiology, Am Mühlenberg 1, 14476 Potsdam, Germany 3- Universidad de los Andes, Biological Sciences Department, Carrera 1 No 18A-12, Bogotá D.C., Colombia

E-62. Prediction of the stage of embryonic stem cells differentiation from genome-wide expression data

Zagar L (1,), Mulas F (2), Sacchi L (3), Garagna S (4), Zuccotti M (5), Bellazzi R (3), Zupan B (1,6,2)*

The developmental stage of a cell can be determined by cellular morphology or various other observable indicators. Such classical markers can be complemented with modern surrogates, like whole-genome transcription profiles, that can encode the state of the entire organism and provide increased quantitative resolution. We hypothesize that such profiles provide sufficient information to reliably predict cell's developmental stage.

Materials and Methods

The test data we have considered includes microarray measurements for a small number of samples from different developmental stages. We start by selecting the most informative genes to reduce the computational costs and remove genes with close-to-constant or too noisy expression levels. Using unsupervised (PCA) or supervised (PLS) data mining methods we then construct prediction models that can predict the differentiation stage of a new sample. A variant of leave-two-out testing was used to evaluate the approach on data with known staging.

Results

In a series of experiments comprising 14 data sets from the Gene Expression Omnibus we demonstrated that the selected approaches are robust and have excellent prediction ability, both within a specific cell line and across different cell lines. Performance was measured using the AUC score and was very high for both cases. Most AUCs were above 0.8 (very good prediction), and quite a few above 0.9 (excellent prediction).

Discussion

Developmental biology is in need of techniques that would accurately assess the progression of cells through development, and predict the developmental stages of cells observed under different physiological conditions. We have presented a computational technique that can provide accurate predictions, even when the prediction model has been trained on cell lines different from those for which the staging needs to be assessed.

Presenting author

Lan Zagar (lan.zagar@fri.uni-lj.si)
University of Ljubljana

Author Affiliations

(1) Faculty of Computer and Information Science, University of Ljubljana, Slovenia; (2) Centro Interdipartimentale di Ingegneria dei Tessuti, University of Pavia, Italy; (3) Dipartimento di Informatica e Sistemistica, University of Pavia, Italy; (4) Dipartimento di Biologia Animale, Laboratorio di Biologia dello Sviluppo, University of Pavia, Italy; (5) Sezione di Istologia ed Embriologia, Dipartimento di Medicina Sperimentale, University of Parma, Italy; (6) Department of Molecular and Human Genetics, Baylor College of Medicine, Houston, Texas, USA

Acknowledgements

This work was supported by the Fondazione Cariplo Project “Bioinformatics for Tissue Engineering: Creation of an International Research Group”, by the FIRB project ITALBIONET, and by the grants from the Slovenian Research Agency (P2-0209, J2-9699, L2-1112).

E-63. High throughput sequencing of the Anopheles transcriptome through sexual development

*Koscielny G (1, *), Nolan T (2), Severgnini M (3), Rizzi E (3), Lawson D (1), Kersey P (1), De Bellis G (3), Crisanti A (2)*

Anopheline mosquito species are vectors for the human malaria affecting hundreds of millions of people per year. Understanding mosquito sexual development and fertility is a crucial aspect for implementing novel and more effective strategies to tackle malaria. The molecular mechanisms of sex differentiation in Anopheles gambiae are yet to be elucilated. Understanding the process of sexual differentiation could provide information on candidate genes and sex-specific splicing patterns for inducing selective male sterility in transgenic lines or for sex-controlled expression of lethal genes.

Materials and Methods

Male and female A. gambiae mosquito larvae were separated based on the inheritance of a sex-linked GFP reporter. Total RNA was extracted from male/female larvae in L1 through L4 developmental stages. Full length mRNAs were converted into cDNA and amplified. cDNA samples were sheared into fragments of less than 400bp and sequenced in a 454 run by using a GS-FLX sequencer. Reads were aligned to the A. gambiae genome using exonerate and stored in an Ensembl database. Splice junctions were identified. Sex-specific splicing patterns per developmental stage were classified using statistical tests.

Results

Exon-exon junctions in each stage were identified, quantified from the number of reads and scored using PWMs derived from canonical 5' and 3' splice sites. These junctions were then compared to identify alternatively spliced genes between male and female using statistical analysis. Sex-specific splicing patterns were classified in categories including exons skipping/inclusion, alternative 5' or 3' splice sites, and mutually exclusive exons. Novel alternative spliced variants covering existing gene models or located in the intergenic regions of the A. gambiae genome were identified.

Discussion

In this study, we report the computational identification of sex-specific splicing transcripts of the A. gambiae genome in early larvae developmental stages from 454-based RNA-Seq . Transcriptome maps were generated in each developmental stages in male and female and splice junctions identified. Sex-specific transcripts detected by the current analysis will be experimentally verified by RT-PCR. At the present time, further analysis of the strength of 5' and 3' splice-sites in alternative splice patterns is still required to complement this analysis and understand sex differentiation process.

Presenting author

Gautier Koscielny (koscieln@ebi.ac.uk)
European Bioinformatics Institute

Author Affiliations

1. European Bioinformatics Institute, Wellcome Trust Genome Campus, Hinxton, Cambridge, UK. 2. Imperial College London, Department of Biological Sciences, Imperial College Road, London SW7 2AZ, UK. 3. Consiglio Nazionale delle Ricerche, Istituto di Tecnologie Biomediche (CNR-ITB), Via F. Cervi 93, I-20090 Segrate, Italy.

Acknowledgements

Bill and Melinda Gates Grand Challenges in Global Health grant. VectorBase project funded by theU.S. National Institute of Allergy and Infectious Diseases (NIH-NIAID)

URL

http://www.ebi.ac.uk

E-64. The Complexity of Gene Expression Dynamics Revealed by Permutation Entropy

Sun X (1,3), Zou Y (2), Nikiforova V (1), Kurths J (2), Walther D (1,)*

High complexity is considered a hallmark of living systems. We aim at studying gene expression time series data from the viewpoint of complexity.

Materials and Methods

We investigate the complexity of temporal gene expression patterns using the concept of Permutation Entropy (PE) first introduced in dynamical systems theory.

Results

Applying the PE complexity metric to abiotic stress response time series data in Arabidopsis thaliana, genes involved in stress response and signaling were found to be associated with the highest complexity. Genes with house-keeping functions exhibited lower PE complexity. High-complexity genes were found to have longer upstream intergenic regions and more cis-regulatory motifs in their promoter regions indicative of a more complex regulatory apparatus needed to orchestrate their expresssion. Evolutionarily old genes were found to be associated with decreased PE.

Discussion

We show that Permutation Entropy is a simple yet robust and powerful approach to identify temporal gene expression profiles of varying complexity that is equally applicable to other types of molecular profile data.

Presenting author

Dirk Walther (walther@mpimp-golm.mpg.de)
Max Planck Institute for Molecular Plant Physiology

Author Affiliations

(1) Max Planck Institute for Molecular Plant Physiology, Potsdam-Golm, Germany (2) Potsdam Institute for Climate Impact Research, Potsdam, Germany (3) present address: Molecular Systems Biology, University of Vienna, Vienna, Austria

URL

http://bioinformatics.mpimp-golm.mpg.de

E-65. In silico study of the regulation of sRNAs in Escherichia coli

Ishchukov I (*), Ryan D, Zarrineh P, Cloots L, Thijs I, Engelen K, Marchal K

In bacteria a large part of the sRNAs regulates gene expression by pairing to mRNAs and affecting the mRNA stability and/or translation. This sRNA network and its impact on bacterial regulation is still unknown: it is not clear how many targets are a

Materials and Methods

In this work we aimed at updating the known Escherichia coli transcriptional network with the sRNA regulatory network. This includes predicting the regulation of the sRNAs and extending sRNA regulons of with novel targets. sRNA regulation was studied using de novo motif detection based on phylogenetic footprinting. Intergenic sequences of orthologous sRNA families in different related gamma proteobacteria were subjected to the PhyloGibbs (v. 1.2). Parameters were optimized as described in Storms et al. For target prediction we combined the intaRNA and targetRNA.

Results

We updated E. coli sRNA interaction network and integrate it with the existing transcriptional network.

Discussion

We further validated the TF-motif assignments in our sRNA dataset by using a strategy based on GO enrichment, assuming that if the target genes of respectively the TF predicted to regulate the corresponding sRNA, and the sRNA itself are involved in similar functions, the assignment of the cognate TF to the sRNA should be more reliable.

Presenting author

Ivan Ishchukov (ivan.ishchkuov@student.kuleuven.be)
K U Leuven

Author Affiliations

K U Leuven

Acknowledgements

OE-grant KULeuven

E-66. Efficient query-based biclustering of gene expression data using probabilistic relational models

Cloots L (1), Zhao H (1),Van den Bulcke T (2), Wu Y (1,), De Smet R (1), Storms V (1), Meysman P (1), Engelen K (1), Marchal K (1)*

Biclustering is an increasingly popular technique to identify gene regulatory modules that are linked to biological processes. From a biologist perspective, one might be interested in knowing which other genes are clustering together in a subset of experimental conditions, with a predefined set of genes that are known to have a common function (query genes). Although some other algorithms also allow for query-based searches, they do not combine the advantages of the query-based search with a model based approach for identifying overlapping biclusters from a possibly noisy set of query genes.

Materials and Methods

We present ProBic, an alternative approach to identify overlapping biclusters in gene expression data using Probabilistic Relational Models, which is able to incorporate biological prior information in the form of a set of query genes. Probabilistic Relational Models were recently developed as an extension of Bayesian networks to the relational domain. An extensive evaluation of the algorithm was performed on biological data to investigate the behavior of the algorithm under various parameter settings and input data.

Results

Applied on an E. coli expression compendium, we show that ProBic identifies biclusters of interest to a biologist from a set of good quality seed genes. Moreover, the bicluster identification is quite robust as the set of query genes can contain several 'noisy' genes that are not part of the bicluster of interest, a situation that often occurs in practice. We also investigated the influence of different parameters and show how they affect various bicluster characteristics, and define a set of default parameters.

Discussion

We present a novel method ProBic, which identifies overlapping biclusters in gene expression data by performing directed queries around genes of interest. ProBic identifies biologically sound biclusters and, bicluster identification is robust as the set of query genes can contain genes that are not part of the bicluster of interest. This framework is based on Probabilic Relational Models, which offer an elegant way for describing a biclustering model that is easily extensible towards integrating additional data sources as well.

Presenting author

Yan Wu (yan.wu@biw.kuleuven.be)
Katholieke Universiteit Leuven

Author Affiliations

1 Microbial and Molecular Systems, Department K.U.Leuven, Kasteelpark Arenberg 20, Leuven, Belgium 2 i-ICT, Universitair Ziekenhuis Antwerpen, Wilrijkstraat 10, Edegem, Belgium

F-28. InSilico DB: an efficient starting point for the analysis of curated human Affymetrix gene expression microarray datasets in GenePattern

Venet D (1,), Coletta A (3,*), Taminau J (4), Steenhoff D (4), Bentabet L (1), Walker N (1), Meganck S (4), Delgado Blanco J (3), de Schaetzen V (3), Savagner F (5), Rousseau F (3), Schy J*

There are ca. 500,000 genomic profiles freely available in Gene Expression Omnibus (GEO). Unfortunately, this information is often in a raw form and requires tedious and error-prone retrieval and compilation before it can be used in visualization and analysis tools e.g., GenePattern, R/Bioconductor or Spotfire. We provide a complete, web-based, easy-to-use solution for finding, combining and exporting uniformly processed and curated human Affymetrix gene expression datasets. The resource is available as a public beta version at: http://insilico.ulb.ac.be

Materials and Methods

Human Affymetrix genome-wide gene expression microarray datasets are extracted from the GEO repository, expert-curated and restructured in an SQL database. A web-interface allows searching, selection, and export of selected datasets. The tool also allows the combination of multiple datasets by providing several normalization and merging options. These can be downloaded in text, R/Bioconductor- and GenePattern-compatible formats. A tight integration with GenePattern was achieved through a direct link between the InSilico DB and any GenePattern server.

Results

InSilico DB contains 176 assays representing more than 25,556 samples originally annotated with ~137 keywords and ~400 values; these are used to build a thesaurus containing ~70 standardised keyword terms and ~180 standardized value terms. Raw gene expression measurements are normalized with several popular algorithms available through R/Bioconductor and made available through the interface. Additionally, three merging methods and customizable export options are available.

Discussion

We argue the InSilico DB grants user-friendly access for researchers to gene-expression datasets by facilitating retrieval and necessary prepossessing. This tool is a step towards making reuse and sharing of genomic datasets a reality for clinical researchers, expanding the possible research questions that can be answered, and providing a means for reproducible genomics research. We believe that this approach is promising to tackle the challenges posed by the rapidly increasing amount and diversity of genome-wide measurements.

Presenting author

Alain Coletta (alaincoletta@gmail.com)
U.L.B

Author Affiliations

(1) IRIDIA-CoDE, Université Libre de Bruxelles - Belgique (2) IRIBHM, Université Libre de Bruxelles - Belgique (3) SWITCH, , Vrije Universiteit Brussel - Belgique (4) COMO , Vrije Universiteit Brussel - Belgique (5) Université Nantes - France

Acknowledgements

The authors are supported by IRSIB through an ICT-Impulse programme of the Brussels-Capital Region 2007- InSilico project and a Spin-Off in Brussels 2009 grant, Enlighten Bioscience project. We are grateful to Michael Reich, and the GenePattern development team, and especially Peter J. Carr for the collaborative development of the data link between InSilico DB and GenePattern. We wish to thank all early beta testers for their useful feedback.

G-72. PopCover – selecting peptides with optimal Population and Pathogen Coverage

Lundegaard C (1,), Buggert M (2), Karlsson AC (2), Lund O (1), Perez C (2), Nielsen M (1)*

The vast genomic variation of pathogens and the diversity of host cellular immune system impose great challenges for the rational selection of T cell epitopes for diagnostic or preventive purposes. As no two pathogen infections are likely to be identical, and no two infected hosts will have identical T cell and MHC binding repertoires, the pool of potential epitopes in an HLA diverse population will exceed many thousands, even in a relatively small pathogen as HIV.

Materials and Methods

DATA: 396 full length HIV genomes with annotated tat, nef, gag and pol proteins covering A(50), B(104) ,C(156), D(40) and AE(46) strains. HLA class II frequencies (HLA Allele frequencies database, http://www.allelefrequencies.net) Select 45 alleles with frequencies > 1% in populations > 2000. 36 HLA-DRB1, HLA-DR3,4,5, and 4 HLA-DQ alleles. METHOD: MHC class II binding predictions by NetMHCIIpan (www.cbs.dtu.dk/services/NetMHCIIpan).

Results

64 peptides were screened in elispot assays against a cohort of 38 patients infected with diverse HIV subtypes, of diverse ethic background and with CD4+ T cell counts > 350. Preliminary results from this large-scale screening shows that between 30% and 100% of the peptides from each HIV protein are recognized by one or more of the patients. 70% of the selected peptide were shown to be immunogenetic, and all 38 patients except one reacted to at least one of the 64 peptides (each patient reacted to on average of 4 peptides).

Discussion

The Popcover strategy for rational selection of peptide pools with broad pathogen and HLA validation has been shown powerful. Preliminary immuno-assay screening indicates that the peptides provide close to complete population coverage. 70% of the selected peptide were shown to be immunogenetic in one or more of the 64 patients, and 37 of the 38 patient (97%) reacted to one of more of the selected peptides. The PopCover method thus shows the potential to provide tailor made peptide pools to cover specific population without prior individual HLA typing and without the use of HLA supertypes.

Presenting author

Claus Lundegaard (lunde@cbs.dtu.dk)
Technical University of Denmark - DTU

Author Affiliations

(1) Center for Biological Sequence Analysis, DTU, Lyngby, Denmark. (2) Department of Microbiology, Cell Biology, and Tumor Biology, Karolinska Institutet, and The Swedish Institute of Infectious Disease Control, Stockholm, Sweden.

G-73. A framework for functional selection of biomarkers

Kivinen V (1,), Nykter M (1), Yli-Harja O (1), Shmulevich I (2)*

Molecular biomarkers measured from blood can be used for evaluating the physiological state of an individual. A well chosen set of biomarkers can be used for disease diagnosis or prognosis, or for monitoring other health conditions. However, identification of such a set is a challenge. Despite the rapid development of measurement techniques the number of molecules that can be measured is limited, especially when considering blood secreted proteins. Thus, for clinical applications the size of the biomarker set should be kept small to minimize the cost and time required for the measurements.

Materials and Methods

We have developed a framework for molecular biomarker selection relying on functional annotations of genes, proteins, or other biomolecules of interest. Annotations are represented as a graph where the functional categories such as pathways and their member genes are used as a basis for selection. Biomarkers specific to a few categories are preferred, though the overall goal is to cover as many relevant categories as possible. We have shown with a mouse gene expression dataset that pathways enriched in the data are covered by our biomarker selection more frequently than other pathways.

Results

In this work, the proposed framework is used for finding disease-specific biomarkers. Here, a bipartite graph between genes and diseases is constructed using a resource such as OMIM for obtaining the connections. The goal is to cover diseases of interest with the selected genes, constituting a collection of diagnostic markers.

Discussion

The presented framework for biomarker selection can be modified to suit different applications aside from the selection of disease-specific markers. A straightforward analysis is to test whether a set of markers is able to discriminate between samples collected from two biological conditions. The selected biomarkers are thus used as features in a classifier. It is also possible to construct a graph specific to a particular condition in a data-driven manner. Here, measurement data can be used for computing association scores to connect genes and pathways.

Presenting author

Virpi Kivinen (virpi.kivinen@tut.fi)
Department of Signal Processing, Tampere University of Technology

Author Affiliations

(1) Department of Signal Processing, Tampere University of Technology, Tampere, Finland (2) Institute for Systems Biology, Seattle, USA

Acknowledgements

This work was supported by the Academy of Finland (project No. 122973, No. 132877, and No. 213462), and the National Technology Agency of Finland (TEKES).

G-74. iPath: Interactive Pathways Explorer

Yamada T (1,), Letunic I (1), Okuda S (2), Bork P (1)*

The KEGG database provides global overview of metabolic pathways. In order to visualize, navigate, explore and analyze the global pathways map, we have developed an interactive pathway explorer.

Materials and Methods

The underlying global pathways map was originally constructed using approximately 120 KEGG pathways, and has been greatly extended in the current version. The map gives an overview of the complete metabolism in biological systems. Nodes in the map correspond to various chemical compounds and edges represent series of enzymatic reactions. In addition, iPath contains a hand-picked selection of important regulatory pathways, extending its usefulness when applied in various metagenomics analyses.

Results

iPath provides powerful tools to visualize, navigate, explore and analyze all, or a subset of, various pathway maps. iPath also offers powerful data mapping tools. Users can upload various types of data to generate custom representations of the map. These customized maps allow users to overview their own data, which is very suitable for genomic/metagenomic projects. Here we show successful use cases. For example, by merging human genome data with metagenomic data from human gut microbiome, iPath clearly shows complementarities of host-symbiont metabolic capacity.

Discussion

iPath is a powerful framework for the exploration and analysis of particular metabolic pathways or overall metabolism, and for comparative analyses of various genomics, transcriptomics or proteomics datasets. The many new display and analysis opportunities provided by iPath include overviews of the metabolic capacity encoded by a single (meta)genome, exploration of metabolic differences in various spatial and temporal series datasets, comparative and evolutionary studies with many organisms in addition to species-complementarity studies.

Presenting author

Takuji Yamada (takuji.yamada@embl.de)
EMBL - Structural and Computational Biology Unit (Bork group)

Author Affiliations

(1) EMBL - Structural and Computational Biology Unit (Bork group) Meyerhofstrasse 1, 69117 Heidelberg, Germany (2) Department of Bioinformatics, College of Life Sciences, Ritsumeikan University 1-1-1 Nojihigashi, Kusatsu, Shiga 525-8577, Japan

G-75. In silico comparative modeling of MTHFR A1298C polymorphism in acute leukemia

*Ramos F (1), Lima J (1), Melo J (1, *)*

Leukemia is the most common malignancy diagnosed in children. Alterations in folate levels induce the development of risk factors for the cancer. MTHFR plays a central role in converting folate to methyl donor for DNA. Recently, the C677T and A1298C mutations of MTHFR were discovered to be associated with susceptibility in acute leukemia. A1298C showed a paradoxal comportment and in vitro studies were not conclusive about the role of this mutation in cancer. These contradictory results have motivated these propose of simulations in silico in order to clarify these results.

Materials and Methods

Initially, this insipient study involves the use of computer tools as the program TRITON, a graphical tool for computational aided protein engineering, for simulate DNA sequence variants with MTHFR, specifically the single nucleotide polymorphism (SNP) A1298C. So, the involvement of this enzyme in the disorders will be study, trying to understand their structural aspects. The interpretation of these in silico results aims to guide new studies about their regulatory interactions and subsequently to propose a qualitative model to analyze the metabolic pathway in silico.

Results

Two of the most investigated single nucleotide polymorphism at MTHFR are C677T and A1298C. The associate of these mutations with decrease susceptibility in acute leukemia is a recent result. But, in vitro experiments with A1298C showed a paradoxal comportment, and so the studies were not conclusive about the role of this mutation in cancer. Our study intends to clarify the contradictory results of case control studies in all continents.

Discussion

This study efforts the use of comparative modeling in order to understand the relationship between sequences challenges and structural consequences. In silico analysis is a trend in wide rage applications to proteomics and it is being used with the purpose of to understand a paradoxal comportment in a SNP at MTHFR. The results of this work will guide new research lines involved regulatory interactions and metabolic pathway of this enzyme.

Presenting author

Jeane C. B. Melo (jeane.ufrpe@gmail.com)
Federal Rural University of Pernambuco

Author Affiliations

(1) Statistics and Informatics Department - Federal Rural University of Pernambuco

URL

http://www.ufrpe.br

G-76. Feasibility space as a tool to understand regulation of metabolic networks

Nikerel IE (1,3), Hu F (1), Berkhout J (2,3), Teusink B (2,3), Reinders MJT (1,3), De Ridder D (1,3,)*

We would like to understand how metabolism is regulated, by analyzing the interplay between various levels of cellular organisation. Hierarchical control analysis shows that metabolism is regulated via a nonobvious combination of hierarchical and metabolic regulation. There is a lack of understanding why the cell uses what form of regulation to achieve certain objectives. Such understanding is invaluable, not only to gain a quantitative understanding of metabolic regulation but also to support effective metabolic engineering and synthetic biology.

Materials and Methods

We explore metabolic regulation by model inversion, i.e. by Monte-Carlo sampling enzyme levels and evaluating whether resulting model states are feasible in terms of a number of predefined physical, biological and evolutionary constraints that generally apply to cellular states and state changes. We postulate that the cell can only explore a subpart of the enzyme-metabolite space, which we call "feasibility space". This space is inspected together with the results of hierarchical regulation analysis to find which regulatory states are necessary to achieve feasible states.

Results

We present an initial attempt to construct and visualise feasibility spaces. We illustrate the proposed methodology on a model of glycolysis in S.cerevisiae, with a focus on the storage metabolism. By analyzing feasibility space, we explore the so-called "turbo design" of glycolysis and the regulatory design required to avoid this unstable state.

Discussion

The information on feasibility spaces can be used to study how physiological constraints govern the modes of regulation (hierarchical vs. metabolic) and to derive design principles for regulatory effects in response to a collection of perturbations. This will aid in further experimental design in systems biology, to discriminate among alternative hypotheses, but also to facilitate engineering the biological systems for improved performance.

Presenting author

Dick de Ridder (d.deridder@tudelft.nl)
Delft Bioinformatics Lab, Delft University of Technology

Author Affiliations

(1) Delft Bioinformatics Lab, Faculty of Electrical Engineering, Mathematics & Computer Science, Delft University of Technology, The Netherlands (2) Systems Bioinformatics/IBIVU, Faculty of Earth & Life Sciences, Vrije Universiteit Amsterdam, The Netherlands (3) Kluyver Centre for Genomics of Industrial Fermentation, Delft, The Netherlands

Acknowledgements

This project was funded by the Netherlands Organisation for Scientific Research (NWO) in the Computational Life Sciences programme and forms part of the Kluyver Centre for Genomics of Industrial Fermentation, a subsidiary of the Netherlands Genomics Initiative.

URL

http://bioinformatics.tudelft.nl/

H-17. Gene dosage balance and the evolution of protein interaction networks

D'Antonio M (1,), Ciccarelli FD (1)*

Recent studies on yeast have shown that essential genes, genes coding for members of protein complexes and genes coding for hubs avoid duplication in order to maintain a tight control of the dosage balance. Similar mechanisms of dosage regulation are maintained also in other species, but not in vertebrates, where more complex relationships between duplicability, essentiality and connectivity have been found. The control of gene dosage has therefore changed through evolution, maybe to adapt to the increase in complexity of vertebrates.

Materials and Methods

To understand the mechanisms of the change in gene dosage balance during evolution, we compare gene and network properties in four model species: human, fly, yeast, and E. coli. These species well represent major evolutionary transitions from prokaryotes to higher eukaryotes and display high quality genomic and protein interaction network (PIN) data. We compare origin, conservation and duplicability of genes in each species with connectivity and centrality of the corresponding proteins.

Results

The origin of a gene and its conservation in evolution are closely related to the PIN properties of their encoded protein. A core of ancestral singleton and central hubs is conserved in all four PINs. Vertebrates display a novel group of hubs that appeared with metazoans, duplicated later in evolution, and are involved in complex cellular functions. The dosage balance of this group of hubs is regulated by mechanisms that are alternative to the retention of the singleton status. These genes duplicated via whole genome duplication, are targets of microRNAs and have tissue-specific expression.

Discussion

This study offers novel insights into the evolution of protein interaction networks. In all species, the PIN core consists of ancient and conserved proteins, which do not undergo further adaptations, while vertebrates have adapted to the progressive increase in morphological complexity, developing new mechanisms to control the dosage balance of recent genes. Despite being mostly duplicated, these genes remain sensitive to dosage modifications and constitute fragile points in the vertebrate PIN. This contributes to explain the occurrence of cancer as perturbation of a fragile portion the PIN.

Presenting author

Matteo D'Antonio (matteo.dantonio@ifom-ieo-campus.it)
Department of Experimental Oncology, European Institute of Oncology, IFOM-IEO Campus

Author Affiliations

1 Department of Experimental Oncology, European Institute of Oncology, IFOM-IEO Campus, Via Adamello 16, 20139 Milan, Italy.

Acknowledgements

The authors thank Vera Pendino (IEO, Milan) for the help in the analysis of miRNA targets, the members of the Ciccarelli lab for useful discussions and the European School of Molecular Medicine (SEMM) for the support. The work is supported by the Start-Up grant of the Italian Association for Cancer Research (AIRC) and by the Fondazione Cariplo to FDC.

I-41. Detecting human proteins involved in virus infection by observing the clustering of infected cells in siRNA screening images

Suratanee A (1,2,), Rebhan I (3), Matula P (1,2), Kumar A (3), Kaderali L (4), Rohr K (1,2) , Bartenschlager R (3), Eils R (1,2), König R (1,2)*

Detecting human proteins that are involved in virus entry and replication is facilitated by modern high-throughput RNAi screening technology. However, hit lists from different laboratories have shown only little consistency. This may be caused not only by experimental discrepancies, but also by not fully explored possibilities of the data analysis. We wanted to improve reliability of such an analysis by combining a population analysis of infected cells with an established dye intensity readout.

Materials and Methods

Viral infection is mainly spread by cell-cell contacts. As a consequence of viral cell-cell spreading, clusters of cells may be formed. Images of knocked down cells for 719 human kinase genes were analyzed. We applied an image processing methods segmenting and identifying infected cells in siRNA screen images. A statistical clustering method (K-function) was employed to define knockdowns harming viral replication in which virally infected cells didn't show any clustering and therefore were hindered to spread their infection to their neighboring cells.

Results

A statistical clustering method was employed and yielded 30 promising candidates suiting as potential host factors for therapeutical drug targeting. The results were compared with a common intensity readout of the GFP expressing viruses and a luciferase based secondary screen yielding five promising host factors which may suit as potential targets for drug therapy. They are comprising CD81, PI4KA, CSNK2A1, SLAMF6 and FLT4. The results are compared to results from the Dengue virus (DV).

Discussion

We report of an alternative method for high-throughput imaging methods to detect host factors being relevant for the infection efficiency of viruses. Applying a clustering analysis method for estimating the virulence in cellular assays is new and can be used for other screens to observe infectious propagation in cellular populations.

Presenting author

Apichat Suratanee (a.suratanee@dkfz.de)
Department of Bioinformatics and Functional Genomics, IPMB, BioQuant, University of Heidelberg, and B080, German Cancer Recearch Center, Germany

Author Affiliations

(1) Department of Bioinformatics and Functional Genomics, Institute of Pharmacy and Molecular Biotechnology, Bioquant, University of Heidelberg, INF 267, (2) Division of Theoretical Bioinformatics, German Cancer Research Center (DKFZ), INF 280, (3) The Department for Infectious Diseases, Molecular Virology, University of Heidelberg, INF 345 and (4) Viroquant Research Group Modeling, Bioquant, University of Heidelberg, INF 267, 69120 Heidelberg, Germany.

Acknowledgements

BMBF-FORSYS Consortium, Viroquant (#0313923); the Landesstiftung Baden-Württemberg (research program RNS/RNAi, contract no. P-LS-RNS30); the Helmholtz Alliance on Systems Biology of Signaling in Cancer; the Nationales Genom-Forschungs-Netz (NGFN+) for the neuroblastoma project ENGINE; the Deutscher Akademischer Auslandsdienst (DAAD); Travel fellowship supported by ECCB10.

I-42. A benchmark on cancer classification using LS-SVMs and microarray data

Popovic D (1,), Daemen A (1), De Moor B (1)*

Because methodological choices at multiple steps in the model building process for classification with microarray data can influence performance, a benchmarking study was performed on two cancer data sets (prostate and breast). Different settings for preprocessing, feature selection, the kernel function and kernel parameter optimization scheme were considered for the Least Squares Support Vector MachineLS-SVM classifier.

Materials and Methods

For the mapping from probe sets to genes, the default CDF files of Affymetrix were compared to customized CDFs (Dai et al, 2005). As preprocessing methods RMA and MAS5 were considered, whilst five feature selection methods were compared (T-test, F-test with outlier removal, F-test on Q values, permutation-based test, DEDS). For the kernel method, the linear and radial basis function (RBF) kernel functions were chosen. Finally for the optimization of the kernel parameters, the influence of the use of a grid search (two types) or and Bayesian framework on performance was tested.

Results

None of the applied methods in each model building stage showed a significant advantage over the others for most of the cases. However,RMA performed better than MAS5 on the prostate cancer data set, and the linear kernel performed better than RBF on both data sets. The choice of the kernel parameter optimization technique, in particular of its cost function, crucially influenced the quality of the derived classifier. The highest classification accuracy and AUC were obtained when Bayesian tuning was used, followed by that by grid search with balanced error.

Discussion

Further validation is needed as observed effects are too weak to draw solid conclusions. We hypothesize that RMA behaves better due to the high false positive rate of MAS5, feeding noisy data to the subsequent steps. The improved results obtained with the linear kernel can probably be attributed to the complexity of the problem at hand and the small number of samples. However, in regard to kernel optimization methods, we have shown that the use of cost functions that explicitly or implicitly take into account the imbalance in the data lead to better classifiers.

Presenting author

Dusan Popovic (Dusan.Popovic@esat.kuleuven.be)
Department of Electrical Engineering (ESAT-SCD/SISTA/BIOI), Katholieke Universiteit Leuven

Author Affiliations

1) Department of Electrical Engineering (ESAT-SCD/SISTA/BIOI), Katholieke Universiteit Leuven

I-43. MiRPara: a SVM-based Software Tool for Prediction of Mature MicroRNAs

*Wu Y, Liu H, Rayner S**

MicroRNAs (miRNAs) are ~22-nucleotide small RNAs that can negatively or positively regulate gene expression at the post-transcriptional level. Experimental detection is biased towards highly or ubiquitously expressed miRNAs and computational approaches are therefore necessary. Current software packages are very successful but restricted to a limited range of species or smaller numbers of shorter sequences. The current release of miRBase contains 14197 entries, and there a need for tools that can analyze large numbers of full length genomic sequences from a broad range of species.

Materials and Methods

we use an SVM for prediction, and unlike other methods, instead of defining a broad range of general parameters, we seleted parameters based on results from experimental studies that identified specific features that were important for miRNA processing. We trained three separate models for Animal, Plant & Virus sequences and applied parameter filtering to determine whether an optimum subset existed for each model that could maintain accuracy but reduce the runtime

Results

the accuracy varied with the models. The best result were obtained with Animal sequences with up to 95% accuracy under specific conditions. However, differences in our predicted secondary structures and those in miRBase reduced the accuracy to 80% when performing full length genome analysis. For Virus and Plant sequences only 60% accuracy was achieved, but this was well in excesss of the accuracy achieved when we tested other available software. Parameter filtering not only increased the speed, it also improved the accuracy of our software.

Discussion

There are many miRNA prediction tools available, but generally they are successful when run on a limited set of sequences, and many of them are limited to small numbers of shorter sequences. When tested against these packages, our software was able to analyze full length genome sequence from a wide range of species with greater accuracy. We are now in the process of investigating our false negative data as this appears to identify miRNAs with unusual properties. We are also attempting to train additional plant models to improve our accuracy for these sequences.

Presenting author

simon rayner (simon.rayner.cn@gmail.com)
wuhan institute of virology

Author Affiliations

Bioinformatics Group, State Key Laboratory of Virology, Wuhan Institute of Virology, Chinese Academy of Science, Wuhan, 430071, P.R. of China

I-44. A Platform for Identifying Prostate-Cancer-related MicroRNA and mRNA using the Empirical Bayes Method in analysing Microarray Data

Chen S-T (1), Ng K-L (1,)*

MicroRNAs (miRNAs) are a class of small non-coding RNAs that bind to their target mRNA sequence and induce either translation repression or mRNA degradation. Recent studies have indicated that miRNA may play an oncogenic or tumor suppressor role in tumorigenesis. We investigate the possibility that miRNA can act as an oncogene (OCG) or tumor suppressor gene (TSG). The Empirical Bayes method, an effective framework for studying the relative changes in gene expression, is employed to identify differentially expressed genes (DEGs) in microarray analysis.

Materials and Methods

Prostate cancer microarray data is used as an input for illustration. DEGs are identified using the Empirical Bayes package provided by Bioconductor. Statistical p-value is supplied by Bioconductor for estimating the significance of a DEG. Two publicly available databases, i.e. ncRNAppi and miR2Disease, which consist of miRNA targets' protein-protein interactions (PPI) and disease-related miRNA data respectively, are utilized to identify miRNA-mRNA relations.

Results

Among the 85 DEGs, eighteen genes, i.e. 18.6%, are found belonging to OCG, TSG or cancer-related genes. The adjusted p-values of these DEGs are less than 0.000019. It is found that among these 18 DEGs, three genes, i.e. FOS, TGFBR2 and AXL, are regulated by miR-101, miR-20a and miR-1, respectively. According to miR2Disease, both of the miR-101-FOS and miR-20a- TGFBR2 regulatory relations are found involving in prostate cancer. miR-1 is also found related with four cancer types. A web based interface is set up for information query.

Discussion

A platform has been set up to study the regulatory role of miRNAs in tumorigenesis. An advanced statistical model, the Empirical Bayes method is used to identify statistically significant DEGs in prostate cancer. Several publicly available databases are utilized to identify miRNA-mRNA relations. Certain putative pairs of miRNA-mRNA are confirmed to be cancer related, hence, the effectiveness of the present approach is demonstrated. The main advantage of the present platform is that all the target genes information and disease records are drawn from experimentally verified records.

Presenting author

Ka-Lok Ng (ppiddi@gmail.com)
Department of Bioinformatics, Asia University

Author Affiliations

Department of Bioinformatics, Asia University

Acknowledgements

The work of Ka-Lok Ng is supported by the National Science Council of Taiwan, under grants NSC 99-2221-E-468-016-MY2.

I-45. PRIDE Inspector: a new tool to browse, visualize and review proteomics data

*Wang R (1, *), Ríos D (1), Reisinger F (1), Vizcaíno J A (1), Hermjakob H (1)*

PRIDE Converter is the preferred tool to convert proteomics data from different sources to PRIDE XML, the format required to perform a submission to PRIDE. However, once the PRIDE XML had been created, it was not possible for submitters to review their own data previous to submission, or to browse and look closely at data already in PRIDE, without using the PRIDE web or BioMart interfaces. In addition, since Quality Control is becoming a more and more important topic, the tool should allow a first analysis on the quality of the data and support community data standards (such as mzML).

Materials and Methods

PRIDE Inspector is an open source application, developed using Java Swing and Java 2D API. The XML handling makes extensive use of the Java Architecture for XML Binding (JAXB) and uses the jmzML library to access mzML files. Database connections and queries are based on JDBC for rapid access. In addition, the visualisation module for all spectra, chromatograms and charts utilizes the jFreeChart library with custom additions for enhanced graphics. The code is freely available at http://code.google.com/p/pride-inspector.

Results

PRIDE Inspector (http://code.google.com/p/pride-inspector) is an open source rich client application for accessing MS proteomics data. It supports the file formats PRIDE XML and mzML and also, it gives direct access to a public PRIDE database instance. Importantly, a statistical view offers a snapshot on data quality. Moreover, it provides APIs/libraries which can be reused independently by the scientific community: the PRIDE JAXB library (for quick reading of PRIDE XML files), and the PRIDE mzGraph Browser library (for the visualization and annotation of spectra and chromatograms).

Discussion

With PRIDE Inspector users can now have a close look at their own data before the actual submission, or access data already in PRIDE for data mining purposes. Highly important for journal reviewers and editors, it facilitates the thorough review of submitted data in the prepublication stage, without the need of the PRIDE web interface. Data can be examined based on different views: metadata, spectra centric, or protein/peptide identification centric views. Apart from its visualization features, the major strength of the tool is the possibility to perform a first assessment on data quality.

Presenting author

Rui Wang (rwang@ebi.ac.uk)
EMBL-EBI, Wellcome Trust Genome Campus, Hinxton, Cambridge, UK.

Author Affiliations

EMBL-EBI, Wellcome Trust Genome Campus, Hinxton, Cambridge, UK.

Acknowledgements

Wellcome Trust [grant number WT085949MA]

I-46. The scientist/staff, project collaboration and content management system (PCCMS)

Kumuthini J (1,), Dominy J (1)*

The CPGR provides a number of wet-lab and dry-lab services and research and development in the omics field, using a variety of technologies. The bureaucratic overhead of managing the different data workflows for all these technologies, often used together in a single project, is onerous. Staff/client interactions occur over a variety of heterogeneous media (phone, email, snail mail, lab book).

Materials and Methods

We use a Linux, Apache, PostgreSQL, PHP stack to implement a web based platform for staff and client collaboration with a focus on the confidentiality and security of client and biological data.

Results

The project collaboration and content management system (PCCMS) has been designed from the ground up to be an extensible system which captures all the omics project meta-data and actual data into a single system facilitating remote collaborative experimental design, staff/client communication, and version controlled documentation and issue tracking. The system is partially automated and able to identify projects that use established workflows and pipelines, from projects requiring additional R&D.

Discussion

The core system will be available at the end of August 2010, and we envisage directly integrating the system with financial/accounting systems and data generating platforms in the near future, as well as validating the system for ISO and FDA approval.

Presenting author

Judit Kumuthini (judit.kumuthini@cpgr.org.za)
CPGR

Author Affiliations

Centre for Proteomics Genomics Research

Acknowledgements

We would like to thank our funding body TIA (Technology Innovation Agency).

URL

`http://www.cpgr.org.za`