An automated approach to the in-silico identification of chimeric mRNAs

Alberti S (1), Trerotola M (1), Emerson A (2), Rossi E (2)

(1) Unit of Cancer Pathology, Center for Excellence in Research on Aging, University "G. D' Annunzio", Via Colle dell' Ara, 66013 Chieti Scalo (Chieti), Italy.
(2) High Performance Systems Division, CINECA, via Magnanelli 6/3, 40033 Casalecchio di Reno (BO), Italy.

Motivation

Chimeric mRNAs from two different genes largely arise by mRNA trans-splicing. mRNA trans-splicing post-transcriptionally joins heterologous mRNAs at canonical exon-exon borders, essentially following the rules of canonical cis-splicing. As the construction of cDNA libraries frequently causes cDNA fusion artefacts, largely because of incorrect ligation of independent cDNAs or of abnormal reverse-transcription, a key issue is how to distinguish between bona fide chimeras and in vitro artefacts. We have developed a bioinformatics retrieval strategy, the In Silico Trans-splicing Retrieval System (ISTReS), in order to distinguish between the two. The ISTRes pipeline consists of the following steps: 1. Map the cDNA databank onto the human genome by Blast analysis, masking human repetitive DNA. 2. Filter the Blast output according to score, match length and percentage identity. 3. Group the query sequence segments (mRNA exons) in longer concatamers, each mapping only onto one chromosome. 4. Check for possible chimeric sequences by comparing concatamers. 5. Remove possible cDNA fusion artefacts (e.g. 'sense/antisense' sequences). 6. Structural analysis of remaining sequences to locate mRNA cleavage or poly-A addition signals to provide further evidence of chimeric joins. The procedure has been successfully validated against a set of known chimeric sequences and has also detected two novel chimeric mRNAs [1]. The authors of this work estimate that about 1% of the hybrid sequences in current mRNA databanks are canonically trans-spliced. The aim of this study was to extend the ISTReS procedure to larger datasets.

Methods

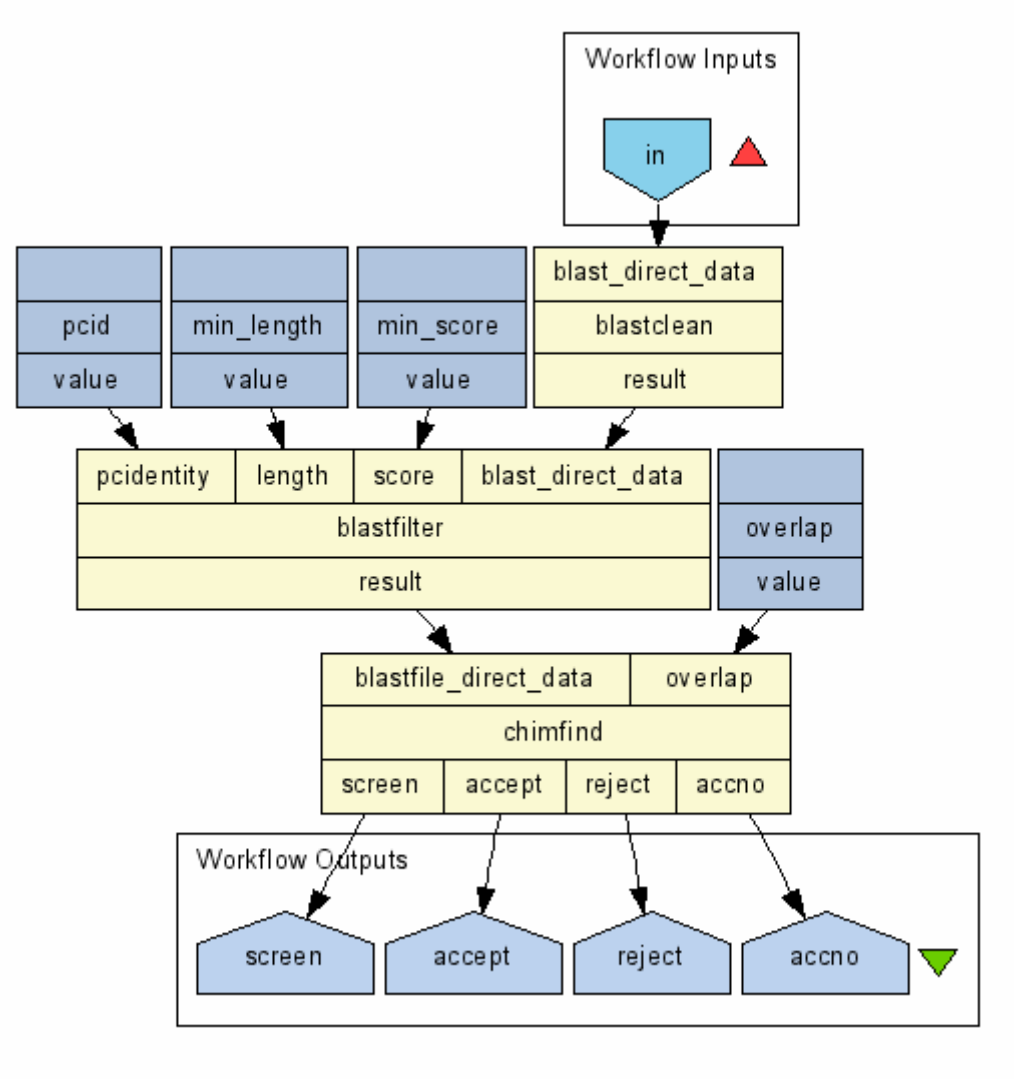
Steps 2-6 of the trans-splicing detection system were implemented with custom Perl scripts, many of them re-written for efficiency and to reflect changes in strategy since the previous study. Although computationally inexpensive the algorithms are often quite complex and most of the programs have undergone major revisions. Indeed, we have found that progress in developing the trans-splicing retrieval system for larger datasets does not depend on the computationally intensive Blast analysis but instead on the validation of the analysis programs. In order to validate the ISTReS procedure the scientific experts in the team need to be able to execute each individual component of the pipeline as well as the whole pipeline itself. The situation is complicated by the requirements for supercomputing resources and large data storage, thus necessitating direct logon access to the computers in question. The common technique of providing a web-interface to hide the underlying computer implementation is in impractical for such a complex system which is still evolving and being tested. To accelerate the refining of the ISTReS procedure and to provide a more convenient environment for the end-user, it was decided to create a workflow description of the pipeline and to implement the various components as web services. The workflow was constructed with the Taverna workflow editor, while the web services were created with the Soaplab environment. The latter is particularly convenient because it generates web services by "wrapping" already existing programs, thereby avoiding re-programming of the applications. Note that due to difficulties in implementing asynchronous web services with available tools, for the moment the Blast analyses have not been exported as web services.

Results

We show below an image of an example Taverna workflow which implements some of the key steps of the ISTReS pipeline. This and similar workflows are currently being used to refine some of

the analysis steps in ISTReS. Candidate chimeras identified by ISTReS analysis with selected cDNA databanks will be reported in a future work.

Contact email: a.emerson@cineca.it



TFBSs prediction by integration of genomic, evolutionary, and gene expression data

Ambesi-Impiombato A (1,2), Bansal M (1,3), Rispoli R (1), Liò P (4), di Bernardo D (1,3)

(1) Telethon Institute of Genetics and Medicine, Tigem, Napoli
(2) Department of Neuroscience, University of Naples "Federico II", Napoli
(3) SEMM, European School of Molecular Medicine, Naples, Italy
(4) Computer Laboratory, Cambridge University, Cambridge, UK

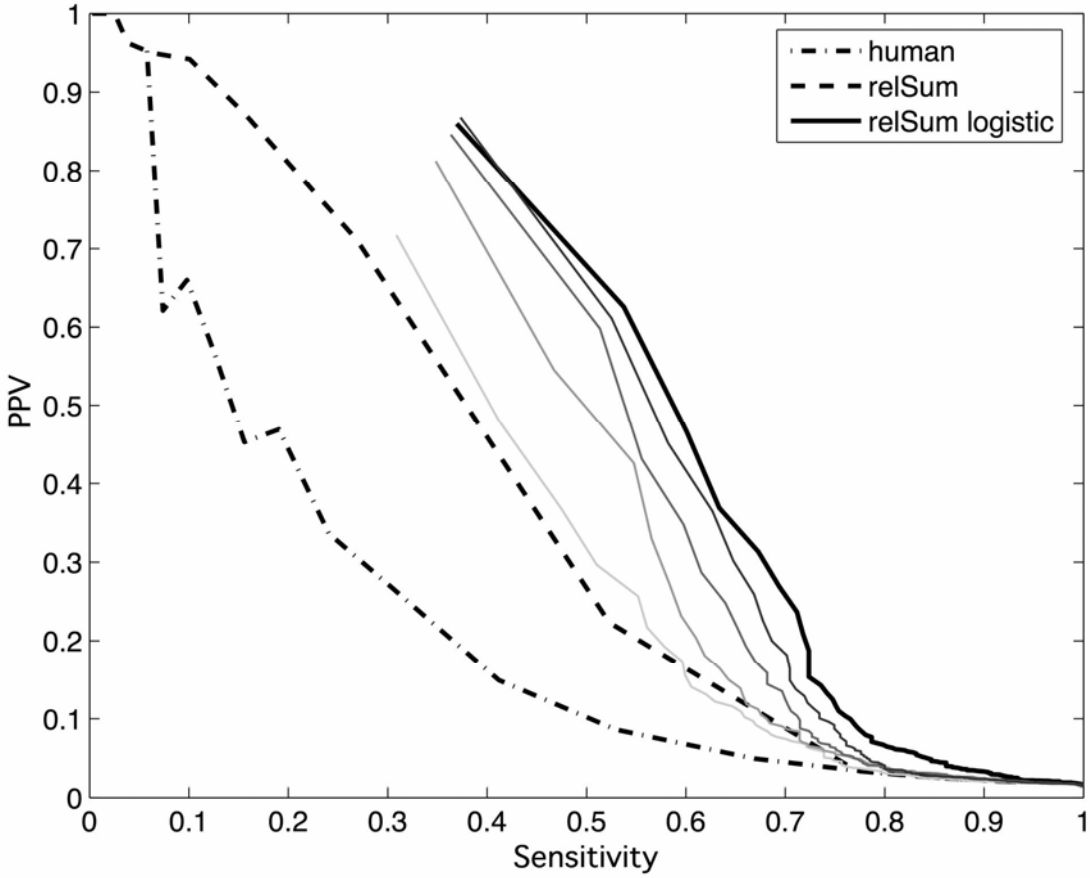
Motivation
Control of gene expression is essential to the establishment and maintenance of all cell types, and is involved in pathogenesis of several diseases. However, biological mechanisms underlying the regulation of gene expression are not completely understood, and predictions via bioinformatics tools are typically poorly specific. We have developed and tested a computatonal workflow to computationally predict Transcription Factor Binding Sites on proximal promoters of vertebrate genes. Finally we applied the workflow to a cluster of genes found to respond significantly to p63 overexpression.This dataset consists of microarray gene expression at 15 time-points in primary murine keratinocytes.

Methods
Our approach for the prediction of regulatory elements is based on a search for known regulatory motifs retrieved from TRANSFAC, on DNA sequences of genes' promoters. Genomic information is retrieved from ensembl database (www.ensembl.org) and compara for orthology information. Predictions are computed independently on different species and the final scores are integrated using a weighted sum calibrated on the phylogenetic distances between the species. These predictions are further refined using logistic regression to integrate data from co-regulated genes. For the purpose of this analysis each matrices were scored using a 3rd order Markov Model trained on a large number of intergenic regions upstream of randomly selected genes.

Results
We show the advantages of integrating genomic data with information based on evolutionary conservation, as well as gene expression data. Consistent results were obtained on a large simulated dataset consisting of 13050 simulated promoter sequences (performance shown in figure 1), on a set of 161 human gene promoters for which binding sites are known. Key factors of our approach include the integration of predictive scores obtained on promoters of ortholog genes from multiple species, and the possibility to include a priori information such as that available from quantitative or qualitative gene expression data, by fitting a logistic regression. A robustness of the logistic regression was evalutated by progressively misassigning genes to the co-regulated group. Our results on simulated datasets show that integrating information from multiple data sources, such as genomic sequence of genes' promoters, conservation over multiple species, and gene expression data, indeed improves the accuracy of computational predictions.

Contact email: ambesi@tigem.it

References
- Tadesse MG, Vannucci M, Lio P: Identification of DNA regulatory motifs using Bayesian variable selection. Bioinformatics 2004, 20:2553-2561.
- Hallikas O, Palin K, Sinjushina N, Rautiainen R, Partanen J, Ukkonen E, Taipale J: Genome-wide prediction of mammalian enhancers based on analysis of transcription-factor binding affinity. Cell 2006, 124:47-59.
- Basso K, Margolin AA, Stolovitzky G, Klein U, Dalla-Favera R, Califano A: Reverse engineering of regulatory networks in human B cells. Nat Genet 2005, 37:382-390.


human
relSum
relSum logistic
PPV
Sensitivity
0
0.1
0.2
0.3
0.4
0.5
0.6
0.7
0.8
0.9
1

Orion: a spatial Multi Agent System framework for Computational Cellular Dynamics of metabolic pathways

Angeletti M (1), Baldoncini A (2), Cannata N (2), Corradini F (2), Culmone R (2), Forcato C (2), Mattioni M (2), Merelli E (2), Piergallini R (2)

(1) Dipartimento Biologia Molecolare, Cellulare ed Animale, Università di Camerino
(2) Dipartimento di Matematica e Informatica, Università di Camerino

Motivation

Computational models that reproduce and predict the detailed behavior of cellular systems form the Holy Grail of systems biology [1]. Molecular Dynamics represents the most accurate and fundamental approach to cell simulation, taking into account the fundamental physical rules at the atomic level. Due to the incredible high number of atoms that must be considered, it cannot be practically used to simulate whole cell systems. A plethora of other mathematical and computational approaches are therefore applied -often experimentally- in systems biology, aiming at the modeling and simulation of cellular systems and processes (e.g. Ordinary Differential Equations, Partial Differential Equations, Petri Nets, UML, PI calculus, Multi Agent Systems, Dynamic Cellular Automata). Methods can be differentiated [2] according to the resolution levels adopted in space, scale and time representation, presence or absence of stochasticity, level of abstraction and to many other factors. The choice of the method implies critical consequences on the model's engineering cycle of life. Issues like accuracy, availability of formal methods to verify properties of the systems, modularity, questions that the model can answer, intuitiveness, scalability, practicability, usefulness for the biological community, existence of suitable experimental data, should all be accurately weighted when choosing a modeling and simulation framework.

Methods

Multiagent systems (MAS) are considered a suitable framework for modeling and engineering complex systems. Agents and agent society permit to intuitively describe a biological system as a set of active computational components interacting in a dynamic and often unpredictable environment [3]. The adoption of such an approach permits to describe the behavior of the individual components and the rules governing their interactions and to observe the emerging behavior of the running system. Our aim is to replicate in-silico the cell system, having the possibility to observe, perturb and inquire a controlled system in order to discover unknown or not too visible correlations between some causes and effects. The attention of systems biology community has been recently attracted by the importance of considering space in the modeling of cellular phenomena. Overcoming the current limitations of Molecular Dynamics, spatial simulation methods should be able to depict coarse grained shapes and size of molecules and their positions in three dimensional space [2]. In [1], Kitano defines Computational Cellular Dynamics (CCD) as the integration of interaction networks approaches with cell-system biophysics. Most interaction-network simulations use the Michaelis-Menten equation or a similar one that assumes a certain ideal condition. However, these assumptions might be unwarranted in a crowded molecular environment in which reactions and molecular movements are constrained in space. In ORION, we therefore propose to provide a physical dimension to the agents representing enzymes and metabolites involved in metabolic pathways. They are collocated and move in a virtual physical 3D space representing the cell cytoplasm. Each metabolic agent acts autonomously and move following Brownian Motion and other laws governing the meso-scale. Enzymes are able to recognize their possible counterparts in known metabolic reactions. The knowledge concerning metabolic pathways is mined from related databases and integrated into a developed domain-specific ontology. When an enzyme agent physically meet, in the space, a substrate agent involved in one of its possible reactions, both undergo to some planned destruction and a new specie of agent arose representing the intermediate complex. According to the related reaction simulation timing, the latter is also in turn subsequently destroyed to give birth to a product and enzyme agents.

Results
We developed an operational computational description of enzyme and metabolites behavior. Issues like movements, interactions and metabolic reactions have been described from the agent point of view. The implementation of the framework on the Hermes agent platform [4] is under development. Another important target of ORION is that of providing, in the context of the LITBIO virtual bioinformatics laboratory (http://www.litbio.org), a workbench to systems biologist on which to engineer, refine and validate models and simulations.

Contact email: nicola.cannata@unicam.it

References
1. Kitano H. Computational cellular dynamics: a network-physics integral. Nature Reviews Molecular Cell Biology. Vol 7, page 163, 2006
2. Takahashi K, Arjunan SN, Tomita M. Space in systems biology of signaling pathways--towards intracellular molecular crowding in silico. FEBS Lett. 579(8):1783-8. 2005.
3. N. Cannata, F. Corradini, E. Merelli, A. Omicini, and A. Ricci. An agent-oriented conceptual framework for systems biology. In T. Comp. Sys. Biology, volume 3737 of LNCS, pages 105-122, 2005.
4. F. Corradini and E. Merelli. Hermes: agent-based middleware for mobile computing. LNCS, 3465:234-270, 2005.

Supplementary informations
This work is supported by the FIRB project Laboratory of Interdisciplinary Technologies in Bioinformatics (LITBIO).

A workflow for an orthology-based prediction of protein-protein interaction

Angeletti M (2), Bartocci E (1), Merelli E (1)

(1) Dipartimento di Informatica e Matematica, Università di Camerino, Camerino (MC)
(2) Dipartimento di Biologia Molecolare, Università di Camerino, Camerino (MC)

Motivation

Many genomes have been completely sequenced. However, detecting and analyzing their protein–protein interactions by experimental methods is not as fast as genome sequencing.These activities are very important to understand properties and mechanisms of cellulars system. The knowledge of the complete picture of the interaction of proteins with their ligands of an organism is defined as the interactome, which can be expressed by a map of (scalefree) interconnecting nodes where the edges represent a value of the interaction strength under particular conditions. Parallel to the application of experimental techniques to the determination of protein interaction networks and protein complexes, the first computational methods, based on sequence and genomic information, have emerged. In our approach we have considered the orthology-based method in which two proteins may interact if each one have at least an ancestor in its phylogenetic tree interacting with at least an ancestor of the other.

Methods

We have designed a workflow taking in input the aminoacid sequences of two proteins A and B. In the first step we search for A and B orthologues separately, using COG database (http://www.ncbi.nlm.nih.gov/COG/new), which contains a classification of proteins, from seven complete genomes and five major phylogenetic lineages, according to their orthologous relationships. Orthologous proteins share a common ancestor and they have been separated by a speciation event. They usually have the same function. We obtain two sets of A- and B-orthologues (AO and BO sets). Then we search for elements of the AO set and BO set which are known to interact.
In our approach we use BIND (http://www.bind.ca), BRITE (http://www.jaist.genome.ad.jp/brite/) and DIP (http://dip.doe-mbi.ucla.edu/) as protein-protein interactions datasets. From the search result in AO and BO subsets whose members are known to interact we obtain the XI and YI subsets. If they are not null, we can conclude that the query A and B proteins should interact. We have implemented this workflow using BioWMS[1].

Results

We are testing experimentally some produced results using this approach. This workflow will be available soon at http://litbio.unicam.it:8080/biowms.

Availability: http://litbio.unicam.it:8080/biowms

Contact email: ezio.bartocci@unicam.it

References

1. Bartocci E., Corradini F., Merelli E., Scortichini L., BioWMS: a web based Workflow Management System for Bioinformatics. Submitted to BITS 2006.

Multiple Single Nucleotide Polymorphisms analysis of candidate genes in Inflammatory Bowel Diseases by using RLS classifiers

Annese V (1), Latiano A (1), Palmieri O (1), D'Addabbo A (2), Maglietta R (2), Liuni S (3), Pesole G (3,4), Ancona N (2)

(1) U.O. Gastroenterologia, Ospedale CSS-IRCCS, San Giovanni Rotondo, Italy
(2) Istituto di Studi sui Sistemi Intelligenti per l'Automazione, CNR, Via Amendola 122/D-I, 70126 Bari, Italy
(3) Istituto di Tecnologie Biomediche-Sezione di Bari,CNR, Via Amendola 122/D, 70126 Bari Italy
(4) Dipartimento di Biochimica e Biologia Molecolare - Università di Bari, Via E. Orabona 4, 70126 Bari, Italy

Motivation
Crohn's disease (CD) and Ulcerative Colitis (UC) are related chronic inflammatory bowel disorders (IBD). The aetiology of IBD is still elusive, but recent studies have suggested that environmental and immunogenetic factors play an important role in the their development. Many susceptibility regions on different chromosomes have been recently pointed out and the correlation between some Single Nucleotide Polymorphisms (SNPs) and IBD has been investigated. In particular, three variants (R702W, G908R, L1007fs) in CARD15 gene on chromosome 16q12 have been associated with CD in Caucasian population. Moreover, some SNPs in DLG5 (chr 10q23) and OCTN1-2 (chr 5q31) genes have also been identified as correlated to IBD. More conflicting data have been reported for the TNF? and MDR1 genes. Since several genes could be involved, each with a limited impact and a possible gene-gene interaction, the traditional methods of analysis such as linkage and association studies could be ineffective. In order to fully understand the relation between many genetic markers and polygenic diseases, new multivariate methods have to be considered. Such methods have to take into account mutual interactions between multiple SNPs in the estimation of the correlation between genotype and phenotype. Statistical learning theory provides valuable methods which are able to jointly consider multiple SNPs and to detect their correlation with pathology. In this paper we quantify the correlation between a set of SNPs and IBD by using the prediction accuracy of classifiers trained on a finite number of examples.

Methods
We have used Regularized Least Square (RLS) classifiers to measure the correlation existing between a set of SNPs and CD or UC. To this end, we have considered 127 CD case, 127 UC cases and 127 healthy controls. We have tested 12 SNPs on 6 different genes: R702W, G908R and L1007fs on CARD15 gene, -857C>T and -308G>A on TNFα gene, C3435T and G2677T/A on MDR1 gene, SLC22A4 and SLC22A5 on OCTN genes, rs124869 on DLG5 gene, and IGR2096Ms1 and IGR2198a on chromosome 5.

Results
We showed that RLS classifiers with second degree polynomial kernels have prediction accuracy of 58.6% in the classification of CD samples versus healthy controls. Moreover, we found a prediction of 50.8%, obtained with linear RLS, in the classification of UC examples versus healthy controls, indicating that the selected SNPs set is more correlated with CD than UC.

Contact email: ancona@ba.issia.cnr.it

Computational detection of cancer-specific splice sites

Anselmo A (1), Iacono M (1), Felice B (2), Guffanti A (2), Pesole G (3)

(1) Department of Biomolecular Sciences and Biotechnologies, University of Milan, Italy
(2) IFOM - The FIRC Institute of Molecular Oncology Foundation, Milan, Italy
(3) Department of Biochemistry and Molecular Biology, University of Bari, Italy

Motivation
Alternative splicing is a mechanism allowing the generation of multiple transcripts from a single gene. This can lead to the expression of structurally and functionally different proteins. Recent studies have shown that alternative splicing can also play an important role in modulating gene expression in different tissues or developmental stages. Moreover, some alternatively spliced isoforms are associated with diseases (different isoforms of some genes such as MLH1, APC and hSNF5 are expressed in normal and neoplastic tissues). For example, mutations in the splice sites of the MLH1 gene induce a double exon skipping that ultimately leads to nonpolyposis colorectal cancer (Venables 2004). Here we present a systematic search of tumour-specific splice sites in order to find new tumor markers.

Methods
We previously developed an algorithm and a web based tool (ASPIC) for the prediction of alternative splicing (Bonizzoni, Rizzi et al. 2005). The algorithm is based on the alignment of multiple transcript sequences against their correspondening genomic sequence. We are now further developing the ASPIC tool by introducing a new module that infers the library source of the ESTs supporting each predicted intron. Knowing the tissue source of each spliced EST, a suitable statistic (with an associated P value) indicates the possible library specificity of splice events (e.g. splice sites or introns). This approach can be used to identify splicing events which are specific to different tissue types or, for example, neoplastic vs "normal" alternative splicing patterns.

Results
We have tested our method on a gene set for which different isoforms are known to be expressed in normal and neoplastic conditions. For these genes, our approach provides good correspondence with experimental results. The systematic application of such method to a larger set of cancer-related genes may lead to the identification of novel "cancer introns", that ultimately can be used to define novel cancer biomarkers.

Availability: The ASPIC software tool can be found in the website http://www.caspur.it/ASPIC/

Contact email: mailto:graziano.pesole@unimi.it

Setting a procedure for "in silico" evaluation of immunoconjugates for cancer therapy

Arcangeli C (1), Gianese G (2), Paparcone R (2), Sperandei M (1), Cantale C (1), Galeffi P (1), Rosato V (2,3)

(1) ENEA BIOTEC-GEN via Anguillarese, 301 Roma
(2) Ylichron srl - Roma c/o ENEA (3)ENEA CAMO

Motivation

A considerable effort is currently applied to the conception and validation of computational methods for predicting biopharmaceutical properties of new drugs. It is expected that they will accelerate the whole process of drug discovery either by the rational optimization of therapeutics candidate or by permitting the rapid exclusion of the poor ones. Therapeutics immunoconjugates consists of a specifically tumor-targeting engineered antibody covalently linked or chelated to a toxic molecule. Their use to selectively deliver drugs to tumor caused great expectations initially disappointed. Recently, the first immunoconjugate has been approved by the Food and Drug Administration (FDA). We have used a computational approaches to investigate some biochemical characteristics of immunoconjugates.

Methods

We used an immunoconjugate consisting of scFv antiHer2 (FRP5) covalently linked to the etoxinA from Pseudomonas aeruginosa as a model system. The tridimensional structure of this immunoconjugate was not available and was obtained by homology modelling. The evaluation of structural characteristics related to biochemical properties will be carried out using molecular dynamics.

Results

We attempt to evaluate some properties as solubility, stability, toxicology of the molecule. We have also used the same procedure to model a second immunoconjugate consisting of scFv antiHer2 (scfv800E6) covalently linked to the same toxin. A comparison between the two models has been carried out.

Contact email: cantale@casaccia.enea.it

Automatic Extraction and Classification of Bioinformatics Publications through a MultiAgent System

Armano G, Manconi A, Vargiu E

Deparment of Electrical and Electronic Engineering, University of Cagliari, Cagliari

Motivation

A growing amounts of information are currently being generated and stored in the World Wide Web (WWW). Digital archives, like PubMed Central, or online journals, like BMC Bioinformatics, are more and more searched for by bioinformatics researchers to download papers relevant to their scientific interests. These services provide search and browsing facilities based on the papers' list of references. However, for a researcher, especially for a beginner, it is still very hard to determine which papers are in fact of-interest without an explicit classification of the relevant topics s/he is involved in. In our view, personalization and effective information-filtering techniques are primary features to be provided. In fact, beyond conventional search engines, users need specific tools and methods for an effective use of all the available scientific resources. In this work we present a multiagent system explicitly devoted to extract information from heterogeneous sources, and classifying them using text categorization techniques.

Methods

Searching for publications involves two main activities: information extraction and text categorization. To this end we devised a multiagent system upon the generic PACMAS architecture. PACMAS, which stands for Personalized, Adaptive, and Cooperative MultiAgent Systems, is a multiagent architecture aimed at retrieving, filtering and managing information among different and heterogeneous information sources. PACMAS agents are autonomous and flexible, and can be personalized, adaptive and cooperative, depending on the given application. The overall architecture encompasses four main levels (i.e., information, filter, task, and interface), each being associated to a specific role. The communication between adjacent levels is achieved through suitable middle agents, which form a corresponding mid-span level. At the information level, agents play the role of wrappers, each one being associated to a different information source. In the current implementation, agents wrap information sources that provide scientific publications; i.e., BMC Bioinformatics site and PubMed web services. Furthermore, an ad-hoc agent wraps a suitable taxonomy extracted from the TAMBIS ontology. At the filter level, a population of agents is devoted to manipulate the information through suitable filtering strategies according to classical text categorization techniques. In particular, a set of filter agents removes all non-informative words such as prepositions, conjunctions, pronouns and very common verbs by using a standard stop-word list. A set of filter agents performs a stemming algorithm to remove the most common morphological and inflexional suffixes from all the words. For each class, a set of filter agents selects the features relevant to the classification task according to the information gain method. After selecting the terms, for each document a feature vector is generated, whose elements are the feature values of each term. The adopted feature value is the TF (Term Frequency) x IDF (Inverse Document Frequency) measure. At the task level, a population of agents has been developed, each of them embedding a classifier. In the current implementation, each agent embed a kNN classifier Task agents have be trained in order to recognize a specific class, each class being an item of the adopted taxonomy. Given a document in the test set, each agent, through its embedded classifier, identify the category of the input document. Task agents are also devoted to measure the classification accuracy according to the confusion matrix. At the interface level, agents are aimed at interacting with the user. Interface agents are also devoted to handle user profile and propagate it through the middle agents. At least in principle an interface agent might also adapt to the changes that occur in the preferences and interests of the corresponding user through a suitable feedback mechanism.

Results

To evaluate the effectiveness of the system, a suitable training dataset is required. To this end, an online support (http://iasc.diee.unica.it/bioclassifier/index.jsp) has been provided to classify and subsequently collect articles according to the adopted taxonomy. Currently, preliminary tests has been performed using a small number of publications classified by an expert of the domain according to the first level of the proposed taxonomy. Let us pointing out that, particular care has been taken in limiting the phenomenon of false negatives (FN), which --nevertheless-- had a limited impact on the percent of false positives (FP). In particular, the ratio FN/(FN+FP) has been kept under 25% by weighting positive prototypes with an additional factor of 1.05 with respect to negative ones. This preliminary experimental results are encouraging and point to the validity of the proposed approach .

Contact email: vargiu@diee.unica.it

A Graphical Tool for Protein Sequences Analysis

Armano G, Saba M, Vargiu E

DIEE - University of Cagliari, Piazza d'Armi, I-09123 Cagliari, Italy

Motivation

Nowadays, one of the most relevant problems in bioinformatics is how to manage the increasing amount of data, empirically produced by researchers involved in life sciences. In fact, in the last few years, an enormous quantity of information has been produced in these fields although the amount of structural information is much more greater than the functional one. To support researchers in discovering the organizational principles and the functional relationships that characterize biological systems, several tools and systems has been devised and implemented. In this paper, a graphical tool for protein sequences analysis is presented, aimed at supporting the user in performing statistical analysis of proteins, in particular at highlighting the relationship between primary and secondary structure.

Methods

Our work is driven by the underlying assumption that the information about a protein sequence can be spread over a set of numerical signals all originating from the given sequence. In this paper we present a graphical tool that allows the user to devise and test (a pipeline of) suitable filters aimed at highlighting specific properties deemed relevant by the user. The results of this filtering activity can be easily plotted by resorting to standard display facilities (charts, histograms, etc..). It is worth pointing out that the proposed tool provides a specific support for studying the relationship between primary and secondary structure in terms of relevant features such as hydrophobicity, aromaticity, dimension, electrical charge, and so on. The system can analyze single proteins or a set of proteins. In the former case, given a protein P, the system supports two functionalities: (i) converting each amino acid in P according to a selected numeric property (e.g. electrical charge), and (ii) filtering P using a window flowing along it (e.g. counting for each amino acid in P the number of Glycines within the current window). Once that P has been converted into a numerical signal, any standard or user-defined filter can be applied to it (e.g. low-pass filtering), possibly giving rise to a pipeline of filters that progressively modify the sequence according to the user needs. The signal obtained after processing P can also be compared to the "profile" of a specific secondary structure, while attempting to highlight a correlation between the two signals. In the latter case, the system supports two functionalities: (a) a given pipelined operation can be applied to all proteins in the given set, possibly yielding a statistics about its capability of predicting a selected secondary structure, and (b) a specific patterns can be searched for, yielding a statistics about its occurrence in the given set of proteins, possibly focusing on its capability of predicting a specific secondary structure. Figure 1.1. shows the graphical interface of the proposed tool. Written in Python, it is completely "open", meaning that additional "ad-hoc" filtering operations can be integrated in a simple way.

Results

The system is currently under testing. Preliminary results highlight that it is particularly suitable for validating or not hypotheses about potential correlations between primary and secondary structure. For the sake of simplicity, we present a case study focusing on the capability of finding a specific pattern: the zinc finger motif. In this case, firstly the user gives as input the set of proteins to be analyzed (in FASTA, PDB, or XML format), and the pattern to be searched for, using regular expressions. Then, the system returns the match set using a textual and/or a graphical format (in this case a chart). Subsequently, among the overall set of proteins belonging to the match set, the user may choose a protein to be further investigated. For instance, the user could be interested in displaying the profile of the selected protein. The chart corresponding to this query is depicted in figure 1.2.

Contact email: vargiu@diee.unica.it

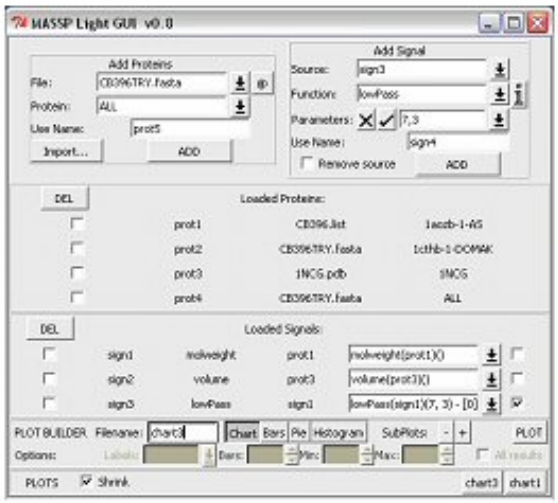


Fig. 1.1 The system graphical interface

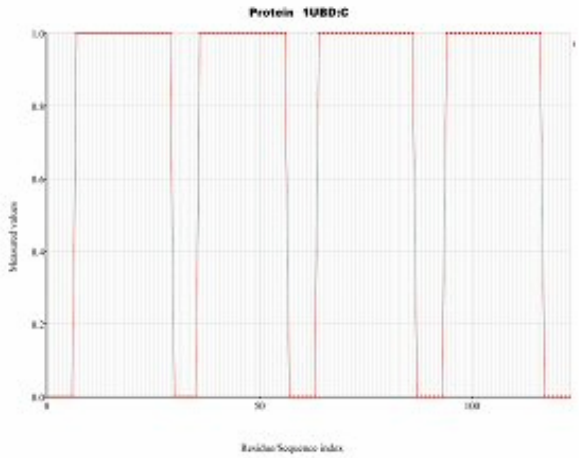


Fig. 1.2 The profile of the 1UBD:C protein

HmtDB, the Human Mitochondrial Genomic Resource: developments in 2006

Attimonelli M, Accetturo M, Jastrzebski JP, Lascaro D, Santamaria M, Zanchetta De Pasquale L

Department of Biochemistry and Molecular Biology , University of Bari, Bari, Italy

Motivation

HmtDB is a Human Mitochondrial genomic resource based on variability studies supporting population genetics and biomedical research (Attimonelli et al.). Here a summary of the most recent improvements as regards both the database and the study of the variability resulting data is reported. In particular, data contained in HmtDB obtained by the application of site variability approaches were studied in order to: contribute in a more rigorous way to a quantitative estimation of the pathogenic proneness of the mutated sites; qualify the importance of the single point mutation in relation to multiple variations associated in one human mtDNA.

Methods

Genomes in HmtDB are analysed in order to estimate their site-specific variability. This is performed on the whole HmtDB dataset and on continent-specific subsets. It has been recently reported in a paper on Human mutation journal, submitted by the present research group and accepted for publication, that variability values can act as haplogroup markers. Further studies on variability data are in progress, in order to estimate the correlation between both nucleotidic and aminoacidic variability and the pathogenic proneness of a specific site. This analysis is also supported by the information available in MITOMAP and by studies of the influence of aminoacidic changes on the secondary and tertiary protein structure. In order to infer a relationship between the variability of a given position and the pathogenic potential of the corresponding mutation, HmtDB data resulting from the application of site variability approaches are compared both to MITOMAP data relative to polymorphic and pathological mutations, and to mtREV index. This last expresses the level of similarity between two aminoacids estimated on homologous multialigned mitochondrial proteins (Adachi and Hasegawa). Moreover, starting from site-specific variability data, changes in the 3D structure and physicochemical features of a modified protein can be investigated. The detection of the influence of single mutation on the whole protein 3D structure is investigated by using such software as Deep View Swiss-PDB Viewer and SWISS-MODEL available at EXPASY, NAMD and VMD, ANTHEPROT, and algorithms developed by the present research group, written in Delphi and C++ language running on the HP ProLiant ML150G2 with OS Linux RedHat and also on PC with WindowsXP.

Results

The current release of HmtDB contains 2155 human mitochondrial genomes of different geographical origin: 200 are from Africa, 660 Asians, 1031 Europeans, 85 from Oceania and 115 Native Americans. Moreover, 500 genomes from subjects affected by "mitochondrial disease" have been collected and will be stored in a separate section of HmtDB database. In 2005, the Query section of HmtDB database was developed by the present research group, and updating and browsing were improved. The Query section allows the extraction, from HmtDB database, of a set of genomes satisfying user-submitted query based on different criteria such as the subject's geographical origin, SNP position, age, sex etc.. As far as it concerns the pathogenic proneness of variated sites, preliminary results about variability, make it possible to postulate the presence of four different classes of variation type: pathological variant associated with disease, with low or average variability value and low mtREV index; rare polymorphic variant, with low variability value and high mtREV index; frequent polymorphic variant, with high variability value and high mtREV index; potential pathological variant, with low variability value and low mtREV index. As far as it concerns 3Dstructure, the calculations and simulations of the changes in the 3D structure of a modified protein suggest that a single aminoacidic substitution can lead to bigger structure

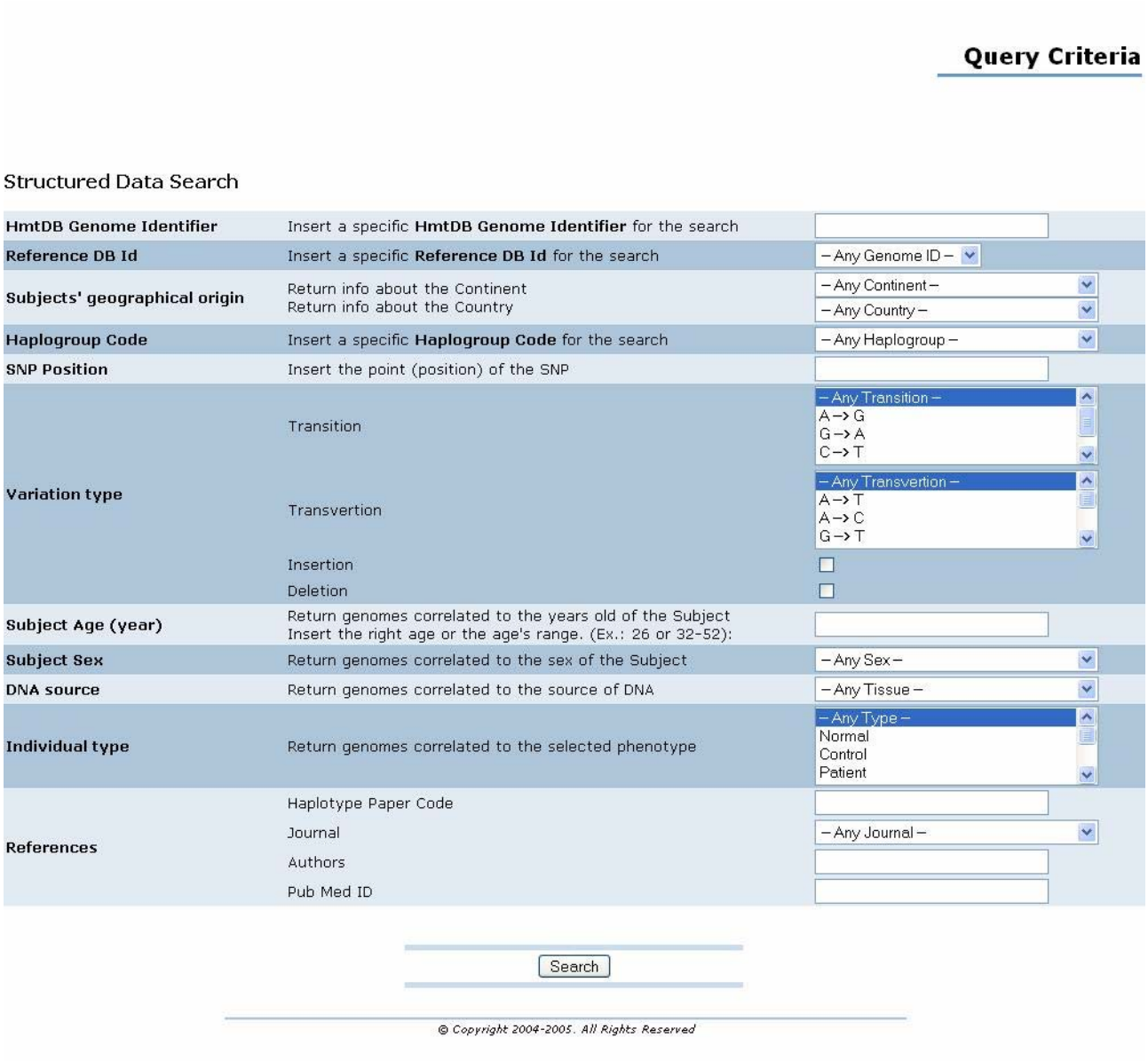
alteration than it appears. Moreover, associations of single variations result in almost twice as big structure changes than the ordinary sum of single punctual mutations. Furthermore, some relatively frequent and seemingly innocuous aminoacidic variations can destabilise the structure of the main part of the protein.

Availability: http://www.hmdb.uniba.it/

Contact email: m.attimonelli@biologia.uniba.it

References

- Attimonelli M, Accetturo M, Santamaria M, Lascaro D, Scioscia G, Pappadà G, Russo L, Zanchetta L, Tommaseo-Ponzetta M "HmtDB, a Human Mitochondrial Genomic Resource Based on Variability Studies Supporting Population Genetics and Biomedical Research". (2005) BMC Bioinformatics, 6(4):S4.
- Adachi J, Hasegawa M. Model of Amino acid Substitution in Proteins Encoded by Mitochondrial DNA. 1996J Mol Evol, 42: 459-468.



A new approach for the analysis of mass spectrometry data for biomarker discovery

Barbarini N, Magni P, Bellazzi R

Department of Computer Science And Systems, University of Pavia, Pavia

Motivation

The recent developments in sample preparation and mass spectrometry allow to measure simultaneously the expression level of thousands of proteins. For this reason, in the last few years an increasing interest has been devoted to the analysis of the body fluids proteome, mainly for diagnostic purposes. In particular the SELDI/MALDI-TOF techniques represent promising tools for the discovery of biomarkers, i.e. the protein signatures associated to a particular disease. However, the identification of such biomarkers is not straightforward due to the presence of several sources of complexity; moreover a well-established procedure for data analysis is not yet available. In the present work, we will propose a new strategy for the analysis of SELDI/MALDI-TOF data based on a three steps procedure for i) data-preprocessing, ii) feature (mass/charge ratio, m/z) reduction and selection and iii) for the association of the selected features to a list of known proteins.

Methods

The proposed methodology for the analysis of the mass spectrometry data consists of three steps. In the first step, many algorithms for the preprocessing of mass spectra are considered. We search for the best sequence of preprocessing algorithms, which is able to maximize the classification accuracy calculated with a simple classifier, based on the difference between the over and under-expressed m/z peaks. The search strategy follows a stepwise approach. In the second step, to decrease the data complexity and to increase the information associated to each feature, the original mass/charge data (e.g. about 300000 in SELDI/TOF high resolution data) are reduced by grouping together the m/z values corresponding to the same protein. Our algorithm exploits the available knowledge on the mass spectrometry technique (e.g. routine resolution) and the chemical properties of proteins (e.g., isotopic distribution). In particular, the algorithm: i) computes the median spectrum of the preprocessed spectra of all subjects; ii) smoothes the median spectrum with a moving window (whose width is equal to the resolution of the spectrometer); iii) finds the maxima of the smoothed curve; iv) computes the sum of the m/z intensities of the isotopic distribution of each maximum (taking into account the routine resolution). The output of this step is therefore a suitable binning of the spectra which is then applied to the data of each subject. In this way, we obtain a reduced number of features that can be used for further analysis. In particular, it is then possible to apply any of the available algorithms to select the most differentially expressed or to define a classifier for diagnostic/prognostic purposes. In the third step, every feature computed and eventually selected in the previous step are associated to a list of proteins that could generate the isotopic distribution. To this aim, a local database composed of proteins and fragments annotated in the Entrez protein database has been created. A list of proteins can be associated to each feature by simply selecting in the local database the entries with molecular weight around the mass of the feature of interest. To reduce the length of such list, it is possible to consider only the proteins that contain at least a peptide discovered in human serum by Plasma Proteome Project.

Results

The proposed methodology has been applied to a public dataset regarding ovarian cancer (http://home.ccr.cancer.gov/ncifdaproteomics/ppatterns.asp). It consists of 216 mass spectra (121 ovarian cancer patients and 95 healthy women) obtained from serum samples by mean of the SELDI-QqTOF technology with WCX2 ProteinChip. The first step of our procedure selected three algorithms for the data preprocessing phase: baseline correction and two smoothing filters (lowess and Sawitzky-Golay). In the second step the initial 373401 m/z were reduced to 3282 features, that correspond to 3282 different isotopic distributions. These features were used as input of several

classification algorithm (e.g, decision trees, k-nearest neighbours, etc.) to discriminate between the two classes (cancer and healthy women). In the third step, the most differentially expressed feature (that classifies the subjects with an accuracy of 90%) was associated to a short list of proteins, among which a possible biomarker for ovarian cancer can be found. The validity of this approach was preliminary tested with success on the protein identification problem of an isotopic distribution reported and experimentally validated in a previous study.

Contact email: nicola.barbarini01@ateneopv.it

A multilayer architecture to support bioinformaticians of today and tomorrow

Bartocci E (1), Cannata N (1), Corradini F (1), Merelli E (1), Milanesi L (2), Romano P (3)

(1) Dipartimento di Matematica e Informatica, Universita' di Camerino, Camerino
(2) Istituto Nazionale di Ricerca sul Cancro, Genova (3) ITB-CNR, Milano

Motivation

In bionformatics fundamental importance are acquiring cyberinfrastructures [1] that will permit multidisciplinary, geographically dispersed, data and computation intensive science. Cyberinfrastructures include peer-to-peer technology, web services and grid technology. In particular grid technology can support virtual communities through sharing of computational and data resource. Simultaneously is growing the request for semantics and the WWW started to become Semantic Web [2]. Nevertheless, scientists difficultly can keep up with the fast development of a specific research area, due to the continuous appearing of new knowledge, data and computational resources. The quest for resources, therefore became a very demanding and time-consuming activity. Bioinformatics deeply changed molecular biology making in-silico experiments a routine task, beside in-vivo and in-vitro ones. In the age of e-Science [3], bioinformaticians can intuitively compose their experiments in the form of workflows. Tasks, designed at a higher conceptual level, are dynamically bound at runtime to physical resources -data and computational ones- taking also into account issues like workload, resource availability and optimization. The integration of all the bio-molecular and "omics" pieces of knowledge requires a significant effort. Built on this premise, systems biology [4] aims at the analysis, modeling and simulation of biological systems and processes, through the supply of mathematical and computational models. Therefore the availability of a virtual desk, on which would be easy to progressively engineer models of biological systems and to simulate and validate them, undoubtedly constitutes another important requirements in modern and future biology.

Methods

To fulfill bioinformaticians needs we propose a multilayer architecture. At the user layer, it is intended to support in-silico experiments, resource discovery and biological systems simulation. The pivot of the architecture is a component called Resourceome [5] which keeps an "alive" index of resources in the bioinformatics domain using a specific ontology of resource information. The Resourceome directly assists scientists in the hard navigation in the ocean of bioinformatics resources. A Workflow Management System, called BioWMS, provides a web-based interface to define in-silico experiments as workflows [6] of complex and primitives activities. In this case high level concepts concerning activities and data could be indexed in the Resourceome. The Resourceome itself would dynamically support workflow enactment, providing the related resources available at runtime. A set of tools for systems biology allows user to intuitively create and refine agent-based models [7] of biological systems and processes. Also in this case Resourcesome can be used to retrieve important related resources like e.g. organism-specific parameters of metabolic pathways. An Agent-based middleware provides the necessary flexibility to support data and computation intensive distributed applications. A middleware permits to develop complex software systems without taking into account at design time who is actually executing them and where they are physically executed. A GRID Infrastructure allows a transparent access to the high performance computing resources required, for example in the biological systems simulation. Beside the computation-intensive aspect, other important issues are taken into account today from grid architectures, like e.g. service grids and knowledge grids.

Results

We conceived the proposed architecture in the context of the MIUR-FIRB LITBIO project(http://www.litbio.org/). The main goals of LITBIO are: to serve the research community with Bioinformatics tools and database and to develop a virtual Laboratory for Interdisciplinary

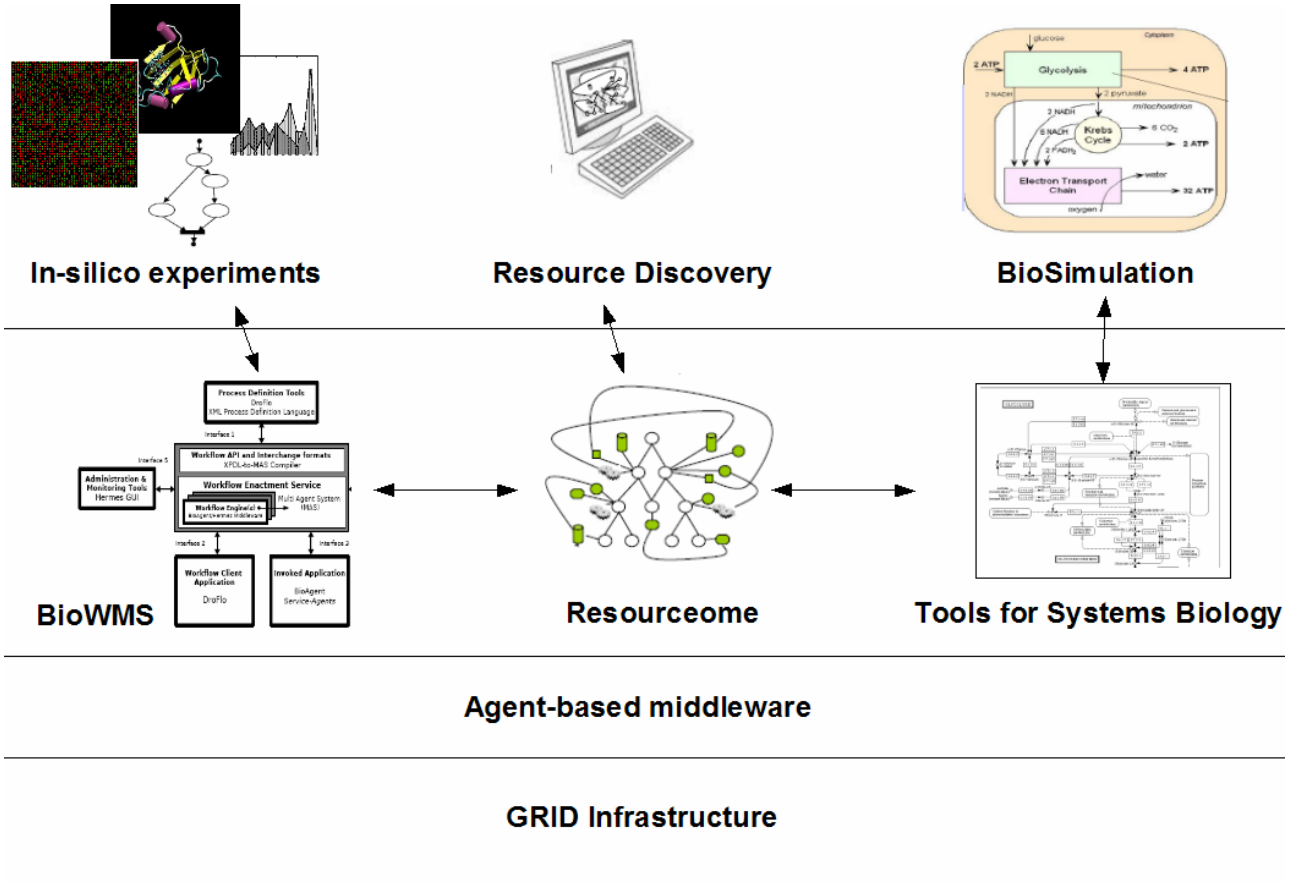
Technologies in Bioinformatics applied to Genomics, Transcriptomics, Proteomics, Systems Biology and Metabolomics.

Availability: http://www.litbio.org/

Contact email: emanuela.merelli@unicam.it

References

1. T. Hey and A. E. Trefethen. Cyberinfrastructure for e-Science. Science, 308(5723):817821, 2005.
2. T. Berners-Lee, J. Hendler, and O. Lassila. The semantic web. Sci Am., 284:3443, 2001.
3. D. De Roure and J. A. Hendler. E-science: The grid and the semantic web. IEEE Intelligent Systems, 19(1):6571, 2004
4. H. Kitano. Foundations of Systems Biology. MIT Press, 2002
5. N. Cannata, E. Merelli, and R. B. Altman. Time to organize the bioinformatics resourceome. PLoS Comput Biol., 1(7):e76, 2005
6. E. Bartocci, F. Corradini, and E. Merelli. Enacting proactive workflow engine in e-Science, Proc. of ICCS 2006, to appear in LNCS
7. N. Cannata, F. Corradini, E. Merelli, A. Omicini, and A. Ricci. An agent-oriented conceptual framework for systems biology. In T. Comp. Sys. Biology, volume 3737 of LNCS, pages 105122, 2005.



BioWMS: a web based Workflow Management System for Bioinformatics

Bartocci E, Corradini F, Merelli E, Scortichini L

Dipartimento di Matematica e Informatica, Università di Camerino, Camerino (MC)

Motivation

An in-silico experiment can be naturally specified as a workflow of activities implementing the data analysis process, in a standardized environment. The workflow owns the advantage to be reproducible, traceable and compositional by reusing other workflows previous defined. A Workflow Management System (WMS), according to Workflow Management Coalition (WfMC) Reference Model [1], is a software component that "defines, manages and executes workflows through the execution of software whose order of execution is driven by a computer representation of the workflow logic". In bioinformatics although several systems that support the daily work of a bioscientists have been proposed in literature [2,3,4], they are not fully compliant to the WfMC standards. In particular, none of them adopts a process definition language standard like XPDL, but each one owns a specific definition language. In many of these systems, workflow editors -i.e. Taverna- are usually embedded with the workflow engine and/or are heavy stand-alone application. A web-based flow designer as WebWFlow, provides a light user-friendly interface to edit, store, share and execute a workflow only using a simple Web browser. Moreover, workflow specifications are generally interpreted and in many cases workflow engines centralize the execution and the coordination of the computation. In the framework of LITBIO[5] project we have developed BioWMS, a web-based WMS, able to dynamically generate domain-specific, agent-based workflow engines from a workflow specification. Our approach exploits the proactiveness and mobility of agent-based technology to embed the application domain features inside agents behaviour. The resulting workflow engine is a multiagent system a distributed, concurrent system-typically open, flexible, and adaptative.

Methods

BioWMS has been implemented on BioAgent/Hermes architecture. Hermes [6] is an agent-based mobile middleware. This choice has been conditioned by the 3-layers user, system, runtime, component-based architecture that facilited the management of domain specific components toward the development of a workflow to multiagent system compiler. User Layer allows bioinformatics to specify their application as a workflow of activities using the graphical notation. In BioWMS we have implemented a web based process definition tool and workflow client application called WebWFlow. System layer provides a context-aware compiler to generate a pool of user mobile agents from a workflow specification. XPDL has been adopted as BioWMS workflow specification language. Run-time layer supports the activation of a set of specialized service agents and it provides all necessary components to allow agent discovery, mobility, creation, communication and security. Service-Agents (SAs) in the run-time layer are localized to one platform to interface with the local execution environment. BioAgent is a tool of specialized cooperative bio-service agents developed to wrapper bioinformatic tools and to perform the primitive activities required by user in an in-silico experiments. User-Agents (UAs) in the system layer are Workflow Executors (WEs), created for a specific goal that, in theory, can be reached in a finite time by interacting with other agents both service and user; afterward the agent will die by killing itself. The Hermes Graphical User Intefrace (GUI) allows the administrator to monitor the status of each platform node, showing the number of WEs running, the SAs activated and memory usage.

Results

As Figure shows, BioWMS, according to WfMC Reference Model, is a WMS that supports the generation of agent based workflow engines.

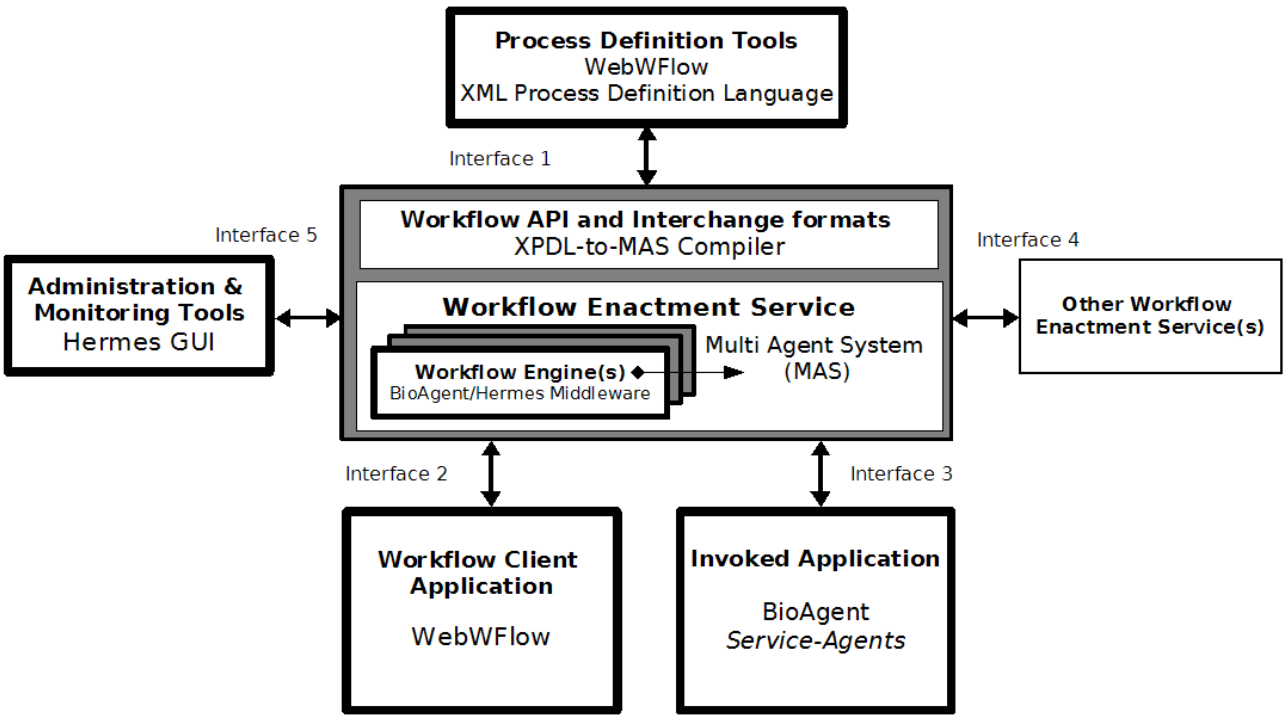
A demo is available at http://litbio.unicam.it:8080/biowms. BioWMS is one of the main components of the LITBIO[6] framework. The aim of this project is to develop a Laboratory for Interdisciplinary Technologies in Bioinformatics and BioWMS helps bioscientist to define in-silico experiments as workflows of complex and primitives activities.

Availability: http://litbio.unicam.it:8080/biowms/

Contact email: ezio.bartocci@unicam.it

References

1. D. Hollingsworth. The Workflow Reference Model, January 1995.
2. T. Oinn et al. Taverna: a tool for the composition and enactment of bioinformatics workflows. Bioinformatics, 20(17):3045-54, 2004.
3. S. Shah et al. Pegasys: software for executing and integrating analyses of biological sequences. Bioinformatics, 1(5):40, 2004.
4. A. Garcia Castro, S. Thoraval, L. Garcia, and R. MA. Workflows in bioinformatics: meta-analysis and prototype implementation of a workflow generator. BMC Bioinformatics, 6(1):87, 2005.
5. Bartocci E., Cannata N., Corradini F., Merelli E., Milanesi L., Romano P. A multilayer architecture to support bioinformaticians of today and tomorrow. BITS 2006. To appear
6. F. Corradini and E. Merelli. Hermes: agent-base middleware for mobile computing. In Mobile Computing, volume 3465, pages 234-270. LNCS, 2005.



A DAS view of agent-based workflows

Bartocci E (1), Merelli E (1), Möller S (2)

(1) Dipartimento di Matematica e Informatica, Università di Camerino, Camerino (MC)
(2) Institut für Neuro- und Bioinformatik, Universität zu Lübeck, Lübeck, Germany

Motivation

The continuous technological advancement in DNA sequencing led to an avalanche of data on biological sequences and their variants. Today, even several large mammalian genomes are fully sequenced and made available i.e. by the Open Source project Ensembl [1]. Sequence annotation could be a very complex task. New sequence features are often inferred after the interpretation, comparison or integration of several data sources and tools. The composition of several activities, described by a workflow, is becoming increasingly important and complex in bioinformatics. The avalanche of data, very often, is too huge to be transferred over the network (e.g. raw images of microarrays). Based on this issue, our approach uses mobile agents as executors of those workflow activities concerning the automation of sequence annotation. A mobile agent is a computational unit capable of migrating to different places from any location. An agent can behave in an opportunistic and reactive way. Agents do not require the user's presence and can be assigned a task to be exploited over distributed resources [2]. In the context of sequence annotation, the agent system should be interactively usable, both by a human and by other programs. This led to the selection of the BioDAS interface [3] for the exchange of biological information. The protocol allows the agent system to disguise itself as a regular DAS server. The data presentation and interactive interpretation is provided by independently developed DAS clients like the Ensembl contig view.

Methods

Our aim was to design an assistant agent -called DAS Interface Service Agent (DASISA) capable to communicate with the external environment using the Distributed Annotation System [4] (DAS) protocol. DASISA to communicate with the DAS protocol, use the Dazzle (http://www.biojava.org/dazzle/) Java Servlet as a general purpose server for DAS protocol. Dazzle is a modular system providing access to several databases. For this work, we have considered an agents' activity results as a data source. Thus, we have developed an AgentDataSource component implementing DazzleDataSource interface. This component communicates through with DASISA exchanging two kind of XML messages: DASAgentRequest and AgentDASResponse. A DAS protocol request is translated by AgentDataSource in a DASAgentRequest document, a message comprehensible by DASISA. The DASAgentRequest document contains the name of the Agent to be created by DASISA and the specific sequence region (segment) id to be explored looking for new possible interesting features. An additional table called "SegmentMap" stores, for each possible segment id, the relative sequence, length and the start position in the inspected chromosome. After having retrieved the necessary segment informations, DASISA creates the agent and sends it a message with the sequence to analyse. Upon termination of the agent's job, its results are returned to DASISA which in turn converts these in an AgentDASResponse XML document. The agents activity may require considerable computational effort. In order to be most responsive for repeated queries, a component called "AgentDASCache" has been implemented to collect the latest results. In the implementation we use BioAgent/Hermes [5] mobile agent-based middleware.

Results

DAS combined with agent technology provides all advantages of dynamic data generation and data integration to otherwise static DAS sources. The approach also addresses the integration of algorithms that may not be feasible to precompute for complete genomes. With the provision of schemas for caching, the individual interests of researchers in a particular family of genes or in a distinct disease-associated locus may be well-addressed.

Contact email: ezio.bartocci@unicam.it

References

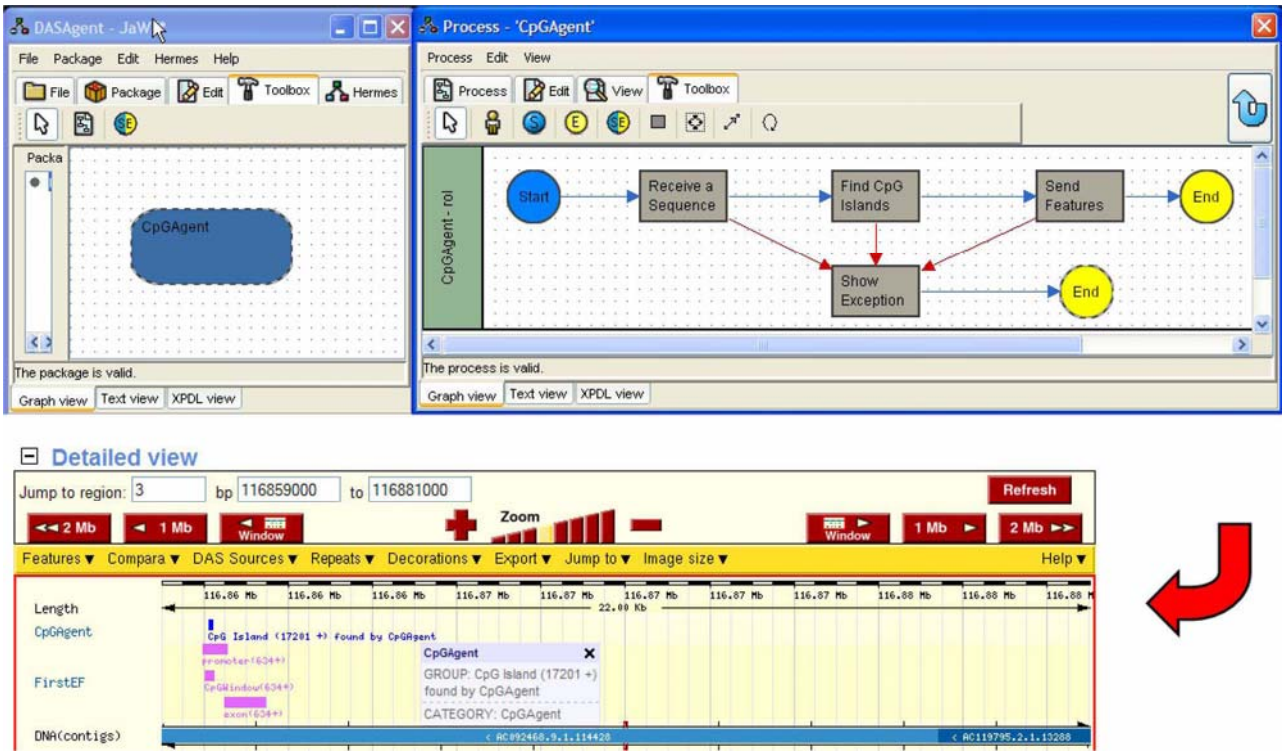
1. Clamp, M., Andrews, D., Barker, D., et al. (2003) Ensembl 2002: accommodating comparative genomics., Nucleic Acids Res., 31(1): 38-42.

2. Hall, D., Miller, J., Arnold, J., Kochut, K., Sheth, A., and Weise, M. (1999) Using Workflow to Build an Information Management System for a Geographically Distributed Genome Sequencing Initiative., Genomics of Plants and Fungi, R.A. Prade and H.J.

3. Bohner, Editors. 4. Stein, L.D., Eddy, S., Dowell, R. Distributed Annotation System., (1999-2002) http://www.biodas.org/documents/spec.html

4. Dowell RD, Jokerst RM, Day A, Eddy SR, Stein L. The distributed annotation system. BMC Bioinformatics. 2001;2:7. Epub 2001 Oct 10.

5. Corradini, F., Merelli, E. (2005A) Hermes: agent-based middleware for mobile computing., Mobile Computing, LNCS, 3465:



A Grid based solution for Management and Analysis of Microarrays in distributed Bone Marrow Stem Cells experiments

Beltrame F (1), Corradi L (1), Milanesi L (2), Papadimitropoulos A (1), Porro I (1), Scaglione S (1), Schenone A (1), Torterolo L (1), Viti F (1)

(1) Department of Computer Science, Control Systems and Telecommunications- DIST-University of Genoa, Italy
(2) Biomedical Technologies Institute (ITB), National Research Council, Segrate, Milano, Italy

Motivation

Exploitation of gene expression data is fully dependent on the availability and sharing of genomic data and advanced statistical analysis tools, which are typically collected on distributed databases/providers and structured under different standards. For these reasons, a Grid based Environment for distributed Microarray data Management and Analysis (GEMMA) is presented. Different microarray (m.a.) analysis algorithms will be offered to the end-user through web interface. A set of independent applications will be published on the portal, and either single algorithms or a combination of them might be invoked by the user, through a workflow strategy. The services will be implemented within an existing grid computing infrastructure to solve problems concerning both the large datasets storage (data intensive problem) and the implied large computational time (computing intensive problem). Moreover, experimental data annotations will be collected according to the same criteria and stored through the Grid portal by using a metadata schema, allowing a comprehensible and replicable sharing of m.a. experiments available in GEMMA among different researchers.

Methods

As first stage a Grid portal will be released, based on Genius. Genius is nowadays a standard of graphical user interface access to the EGEE and Italian Grid infrastructures so it appears as the most suitable and convenient solution in our implementation. To complete this process, applications will be provided as grid services exploiting standard Grid infrastructure (authentication, inter-process communication, data management, job scheduling). From a functional point of view, the adopted framework allows to deploy both servlet components (visible to users as traditional web pages) and services (grid or web-services) exposing key components to the public with standard interfaces. From a data point of view, the proposed environment permits users to upload/download their data and results on/from the Grid Portal and store them on Grid storage resources. GEMMA environment is based on LCG, the official middleware for the Italian Grid infrastructure. A remarkable metadata management system is provided by the ARDA Metadata Catalogue Project (AMGA) metadata management server and clients that can be easily integrated in the LCG environment. The system is directly accessible through the web interface where users fill a multi-page web form with required fields. The form is then processed and submitted to the catalog, as soon as the uploading of data on Grid provides a consistent logical reference to the data themselves. In order to develop a functional distributed management environment, particular attention has been addressed to the description of experiments using standard annotations. MIAME/MAGE guidelines have been adopted, omitting some standard fields and joining others to focus on our specific biological domain. DChip, one of the most complete and diffuse free software for the m.a. data analysis, was chosen for our application to cover tasks from image normalization to chromosome location, passing through filters to eliminate redundant or useless information, sample comparison extracting differentially expressed genes, hierarchical clustering or LDA classification to facilitate bioinformatics analysis. To enhance scalability in our implementation, the use of alternative open source sw for the genomic data analysis and comprehension, such as bioconductor, has also been considered. The proposed environment will be tested by using a specific experimental scenario: human bone marrow stem cells (BMSC) exposed to different experimental conditions.

Results

This platform provides a shared, standardized and reliable storage of biological data related to BMSC culture. Different m.a. analysis algorithms will be offered to the end-user, through a portal web interface, starting from the Linux-ported dChip analysis tool. Several applications can be invoked and combined by the user, through a workflow strategy. Problems concerning large datasets storage, storage safety, and large computational times are solved through a grid computing infrastructure. The Grid portal will act as user interface for data storage, metadata management, data analysis and result retrieval. Data access from processing job will be done directly on distributed file system, without moving m.a. datasets to computing nodes local filesystem. Experimental data annotation will be gathered in GEMMA and stored with the use of a metadata scheme facilitating the apprehension and sharing of m.a. experiments. The tool can be also used to test several algorithms with different parameter configurations on the same dataset simultaneously and it's open to the integration of third-party modules for specific functions (like clustering algorithms and statistical analysis tools). This platform is currently part of the Italian FIRB project LITBIO (Laboratory for Interdisciplinary Technologies in BIOinformatics).

Contact email: pivan@bio.dist.unige.it

Dynamic simulation of protein interaction networks

Bernaschi M (1), Castiglione F (1), Ferranti A (2), Gavrila C (2), Cesareni G (2)

(1) Istituto per le Applicazioni del Calcolo "M. Picone",CNR), Roma.
(2) Department of Biology, University of Tor Vergata, Roma

Motivation

Protein interactions support cell organization and mediate its response to any specific stimulus. A realistic simulation of cell physiology would require (at least) genome wide information about protein concentration combined with integration of a quantitative protein interaction network with the metabolic and gene regulatory networks. These data are currently not available. Nevertheless, simple computational models may help to extract general principles from the available noisy and qualitative information. Protein-protein interaction (PPI) networks, like other kinds of complex networks, are not randomly organized. They display properties that are typical of "hierarchical" networks combining modularity and local clustering to scale free topology. From the analysis of the "static" representation of the corresponding graph, it is not clear why biological networks evolve the observed characteristics. In our project we aim at investigating whether the "static" connectivity properties of a PPI network (scale free, modularity and high local clustering), analyzed in a dynamic model, favor the formation of higher order structures and eventually cell organization. To this end, we designed and implemented a computer model to simulate the interaction of a large number of proteins within a naive unstructured cell (devoid of compartments).

Methods

In order to simplify our cell model and render it computationally tractable we made a number of assumptions. The intracellular space is mapped onto either a two-dimensional or a three dimensional lattice. In the lattice, each site represents an average space of 5 nm, which is comparable with the diameter of an average globular protein. The lattice is filled with proteins with a 20% average occupancy that is compatible with the estimated crowding of proteins in the cell cytoplasm. The simulation is carried out in the following steps: diffusion, rotation, association and dissociation. At each time step proteins in neighboring cells may interact and form a complex depending on the interaction rules (i.e., if there is an edge linking the two proteins in the input PPI network) and on the association constants that define the probability for the proteins to bind and to form a complex. Furthermore, any existing complex can break depending on the dissociation constant. This whole procedure can be seen as a sort of "discrete molecular dynamics" applied to protein interactions in the cell.

Results

We performed different simulations using as interaction rules those derived from the experimental interactomes of E. pilori (724 nodes, 1403 edges), E. coli (1289 nodes, 5420 edges) and S. cerevisiae (1378 nodes, 2451 edges) and we compared their dynamic behaviors with that of random networks having an equivalent number of nodes and edges. The simulations have been done both in the three and two dimensional lattice models. We are currently analyzing the dynamic structures that are formed in natural and random networks to identify the characteristics of the natural networks that favor the formation of an organized virtual cell.

Contact email: massimo.bernaschi@gmail.com

Genome-wide prediction of PCR products based on termodynamic parameters

Boccia A (1), Silvestre A (2), Paolella G (1,3)

(1) CEINGE Biotecnologie Avanzate, Napoli, Italy
(2) Dep. Biochimica e Biotecnologie Mediche, Universita' 'Federico II', Napoli, Italy (3) Dep. SAVA, Universita' del Molise, Campobasso, Italy

Motivation

DNA amplification by polymerase chain reaction (PCR) is one of the most powerful techniques currently used in molecular biology. Nevertheless, the behaviour of the reaction is not completely predictable: non-targeted products are often amplified, particularly when complex templates, such as genomic DNA, are involved in the reaction. Mismatch tolerance is the most significant factor affecting PCR specificity, together with primer length, template size and product size limit. Computer programs are commonly used to accurately design PCR primers on the basis of a range of factors related to primer sequence and to predict the formation of non-targeted products. The specificity of primer/template interaction is tested by comparing oligo and template sequences, but the analysis is typically limited to a restricted region around the target sequence, rather than the real template, often a whole eukariotic genome. More recent programs evaluate primer/template specificity by searching for primer matches in the full genomic sequence, but they usually provide a list of potential amplimers, without attempting a quantitative analysis of amplification products. The tool presented here predicts all potential products, generated by two or more oligos on a given genome. Product yield is simulated by taking into account the stability of primer/template hybrid, calculated by using predicted thermodynamic parameters. The simulation may also be extended to cover RT-PCR by using EST DBs or predicted genes on the genome sequence.

Methods

The procedure was developed by using the PHP programming language and consists of a number of different steps. First, the target genome is searched in order to identify all similarity matches, by using BLAST (7) with modified default parameters to enhance search sensitivity. The search time is significantly reduced by always performing only one BLAST run against a single query sequence formed by concatenating all primers. Matches are then ordered on the basis of chromosome position and orientation, and only those compatible with potential amplification are retained. The stability of primer/template interaction is predicted at these potential priming sites, by calculating free energy, enthalpy, entropy and melting temperature, as described (1-6).

Results

Some commonly used melting temperature prediction methods were evaluated, and a tool to calculate primer/template stability was developed on the basis of the algorithm proposed by Allawi and Santa Lucia (1-6), with minor modifications. The tool was used as a base to build the PCR prediction program described above. The program output returns the predicted priming sites together with the termodynamic parameters. Ratherb than simply determining the expected optimum temperature, the stability of potential hybrids is calculated under the specific conditions used, including temperature, salt concentration and other parameters. The dissociation constant of the primer/template complex is used to quantify the expected yield of each amplimer, estimated as the fraction of template DNA bound to the primer given the primer concentration and the dissociation constant of the primer/template complex. A typical run is complete in 20-30s for the human genome on current hardware, independently of the number of oligos. Experimentally validated cases, where extra bands are produced, confirm that predicted results closely match the observed bands. The prediction may also be carried out for RNA amplification, either through search of EST libraries, or through analysis of transcripts expected according to exon annotation in the genome sequence. Graphical output is provided in the form of a simulated picture of a gel

electrophoresis showing all the expected products and the relative predicted intensities. A web interface is in preparation.

Contact email: boccia@ceinge.unina.it

References

1. John SantaLucia, Jr. et al., Biochemistry 1996, 35, 3555-3562
2. Allawi H.T. et al., Biochemistry,1998, 37, 9435-9444
3. Allawi H.T. et al., Biochemistry, 1998, 37, 2170-2179
4. Allawi H.T. et al., Nucleic Acids Research, 1998, 26 (11), 2694-2701
5. Allawi H.T. et al., Nucleic Acids Research, 1998, 26 (21) 4925-4934
6. Peyret N. et al., Biochemistry, 1999, 38, 3468-3477
7. Altschul, S.F. et al. 1990 J. Mol. Biol. 215:403-410

Multiplatform approach for robust gene identification

Bosotti R (1), Locatelli G (1), Healy S (1), Scacheri E (1), Calogero R (2), Isacchi A (1)

(1) Genomics Unit, Department of Biotechnology, Nerviano Medical Sciences, Nerviano (MI)
(2) Genomics and Bioinformatics Unit, Department of Clinical and Biological Sciences, Az. Ospedaliera S. Luigi, Orbassano (TO)

Motivation
Microarrays have been widely used for the analysis of gene expression and several commercial platforms for measuring genome-wide gene expression levels are currently available.The issue of reproducibility across different platforms has yet to be fully resolved.
Methods

Results
In this paper, we describe a cross platform validation and we suggest that the identification of trustful differentially expressed genes might be improved by cross platform analysis.

Contact email: roberta.bosotti@nervianoms.com

Construction of compositionally biased amino acid substitution matrices to improve annotation of Plasmodium falciparum proteins

Brick K, Pizzi E

Dipartimento di Malattie Infettive, Parassitarie ed Immunomediate, Istituto Superiore di Sanità, Roma

Motivation

Protein alignment algorithms such as BLAST and FASTA, currently use substitution matrices based on average amino acid distributions of conserved domains to score hits. However, due to the method of construction of these matrices, when proteins of biased amino acid distribution, or with large low complexity domains are aligned, the implicit frequencies in these matrices do not reflect accurately the protein composition. Proteins of the most virulant strain of the human malaria parasite, Plasmodium falciparum, exhibit a strong amino acid bias, as a result of a highly AT-biased genome (~80%). Some 60% of the proteins of this organism remain annotated as hypothetical and efforts to align these proteins with those of other known species have met with little success. Our approach aims to improve the annotation of this group of proteins, through the construction and use of matrices which more closely reflect the implicit amino acid bias.

Methods

Amino acid frequency profiles were generated for each of the 28337 blocks of multiple alignments in the BLOCKS database (Henikoff and Henikoff 1991, Smith 1990). A range of substitution matrices were developed using a perl algorithm based on the same underlying mathematical structure as the commonly used BLOSUM and PAM matrices. This model dictates that a matrix can be built in the log-odds form: $sij = 1/\lambda * (qij/pi*pj)$, where there is at least one positive score and the expected score is negative. Our algorithm derived the observed amino acid pair frequencies (qij) and expected amino acid background frequencies (pi) for each set of blocks, created a matrix in half bit units, and provided the statistical parameters (expected score and un-gapped entropy) of each matrix. Each set of blocks was developed into a set of matrices clustered at 100% and 62%. Clustering was based on an original hierarchical clustering method, in which each cluster was composed of sequences all sharing a given percentage identity. When a sequence contributed to more than one block, the contribution of that sequence to amino acid counts was scaled by the inverse of the number of blocks in which it was a member. The alignments yielded by each matrix were then compared with those from a BLOSUM matrix with equivalent gapped statistical parameters. A novel procedure was used to quantitatively compare the statistical parameters (bit score, E-value) and length of each hit to evaluate improvements. This allowed us to select a subset of matrices, (SCAMB) which showed the most improved alignments using this technique.

Results

Our matrices, due their biased and symmetrical nature, are suited to the comparison of proteins with others of similarly biased amino acid distributions. We have shown, that both in searches using BLASTP and the Smith-Waterman algorithm (SSEARCH) a substantial improvement is seen in the alignments of a set of P.falciparum antigen proteins using a subset of these matrices. Specifically, it is shown that annotations of all members of the recently categorised family of parasite erythrocyte membrane proteins, the SURFINS (Winter et al. 2005), are improved when aligned using the SCAMB matrices. 1. PFA0725w is a P.falciparum protein which has been suggested to play a dual role in the life cycle of the parasite, both as an antigen presented on the host erythrocyte membrane and as a protein probably involved in invasion of host red blood cells. When aligned against a protein database, using the SCAMB matrix, similarity to several P.falciparum erythrocyte surface proteins is improved, and novel hits to axoneme proteins from other species are found. These results validate the dual role hypothesis and suggest what the role of this protein in the malarial invasion process may be. 2. Alignments between highly similar members of the same family, show slight changes to the alignments, which result in improved statistical parameters (bit score, E-value). 3. All SURFINS had an increased number of novel hits to long low complexity regions in other

antigens in a P.falciparum protein database. It should be noted that a large proportion of these hits had an increased Bit score, decreased E value and an increased hit length. We suggest that a refinement of this approach may yield further improvements in the annotation of Plasmodium and other biased genomes.

Contact email: kevbrick@gmail.com

References:

- Henikoff,J.G. and Henikoff,S. (1992) Amino acid substitution matrices from protein blocks, Proc. Natl. Acad. Sci. USA, 89,10915-10919
- Smith,H.O. (1990) Finding sequence motifs in groups of functionally related proteins, Proc. Natl. Acad. Sci. USA, 87,826-830
- Winter,G., et al. (2005) SURFIN is a polymorphic antigen expressed on Plasmodium falciparum merozoites and infected erythrocytes., Journal of Experimental Medicine, 11,1853-1863

Analysis of operon genes using a compendium of expression data

Brilli M, Fondi M, Fani R

Dipartimento di Biologia Animale e Genetica dell'Università di Firenze, via Romana 17/19, 50125 Firenze, Italia

Motivation
The origin and evolution of operons is still under debate. One of the most accepted hypothesis concerning the molecular mechanisms and forces responsible for their assembly suggests operons origin and evolution be driven by the co-transcription of the genes they contain, as well as by the spatial co-localization of the products they code for. In principle, the transcription of operons into a single polycistronic mRNA implies an equal transcription levels of all genes. However, it is known that some operons, especially the longest ones, may contain internal promoters responsible for the transcription of distal genes. In this work we tried to assess the degree of correlation existing among the transcription levels of genes belonging to the same operon and to check whether the distance of a gene from the transcription start point might influence the degree of its transcription, by using a compendium of published expression data.

Methods
Expression data were downloaded as supplemental material to published papers, corresponding to 10 different publications and a total of 79 different experimental conditions. Only normalized and filtered datasets were used. All the operons (over 1400 genes) from Escherichia coli, retrieved from the regulonDB website (http://regulondb.ccg.unam.mx/index.html) were used. Specific java classes were written to calculate the Pearson correlation coefficients of expression patterns of the first gene in the operon vs all the downstream genes. A set of 10000 pseudo-randomly generated pair of genes was used as a background. We explored the distributions of correlation coefficients and performed linear regression analyses with the distance of a gene from the transcription start as predictor of the correlation among the expression patterns of that gene and the first of the operon.

Results
We checked our expression compendium to detect the significance of the correlation coefficients calculated for operon genes vs random pairs, by analysing the corresponding distributions. Data obtained revealed that the squared-Rs calculated for operon genes are, on average, greater than the corresponding values calculated for pseudo-randomly generated pairs of genes (Wilcoxon Rank Sum test, p0.72; for the random dataset the value is <50%) confirmed that the operon is very effective in maintaining the coexpression of genes. We then explored the relationship existing between gene location in the operon and the correlation coefficients by performing a linear regression analysis using as predictor variable the distance of a gene from the first of the same operon, and dependent variable the correlation coefficient in expression patterns. A statistically significant negative correlation was found (R-squared=0.087, p<5500 nt in length). Nevertheless, very long operons exist with high expression correlations among all genes, and they do not fit well with the model. For this reason, we extracted information concerning the presence of internal promoters (IP) (source: RegulonDB) and we found a positive correlation among operon size and number of IPs. To check the effect of IPs on the expression levels inside operons, we normalized them with respect to the expression level of the first gene in the operon. This revealed that 62% of the genes in operons without IP have a normalized expression between 0 and 1, suggesting that most of them have a reduced expression with respect to the first gene. On the contrary, only 45% of genes in operons with IPs are in the same range, revealing that internal promoters increase the fraction of operon genes with an expression level greater than that of the first gene. In conclusion, the analysis of the E. coli operon dataset suggested that the size may be a limiting parameter during operon evolution. This can explain why very often operons are extremely compact and that the greater the operon length the greater the number of Ips. Therefore, it is possible that the longest operons might have been assembled by a piece-wise mechanism, according to which they have been constructed by the fusion of shorter pre-existing mini-operons rather than the sequential addition of

single genes, as suggested for the histidine biosynthetic operons (Fani et al. 2005). This example reveals that using compendia of expression data can reveal useful information when large datasets are analysed.

Contact email: r_fani@dbag.unifi.it

Supplementary informations
References to microarray data are available upon request, they were not explicited for shortness.

Network analysis of plasmids encoded proteins

Brilli M, Mengoni A, Fani R

Dipartimento di Biologia Animale e Genetica dell'Università di Firenze, Via Romana 17/19, 50125 Firenze, Italia

Motivation

Plasmids are widespread in prokaryotes and harbour genes coding for a number of activities that can vary between different organisms groups. Plasmids are often transferred among microorganisms, and this permit the spread of new metabolic activities within natural bacterial communities. A clearcut example is the transfer of antibiotic resistances, often coded by plasmid-borne genes. The evolution of plasmids has been studied using classical bioinformatics tools, such as phylogenesis and gene order/function comparisons. However, large scale analyses of plasmid encoded proteins have not been performed. Network analyses have been applied both to protein interaction and gene expression data, and we decided to use networks to represent and analyse the plasmid content and the homology relationships linking the proteins they code for, using dedicated software.

Methods

We downloaded the entire NCBI plasmid encoded protein dataset, and we divided it into sub-datasets in conformity to different grouping rules (e.g. all proteins from Archaea, or those from a single genus or specie). Each subset was used to generate a database by using the formatdb utility of the stand-alone blast program (Altschul et al., 1997), and the very same dataset was then used for an all-against-all blast search. Dedicated java classes were written to: 1. search information concerning plasmid source on the web; 2. calculate coordinates of each node (protein) so that each plasmid is arranged circularly with the encoded proteins on its circonference and separated from other plasmids; 3. read the blast output to write down links among the different nodes, storing the percent identity value corresponding to each alignment (link). The java class writes a file in vsn format, readable by the network representation and analysis software Visone (www.visone.de). The user is prompted at start for E-value and alignment length cut-off values, permitting to explore networks at different detail levels.

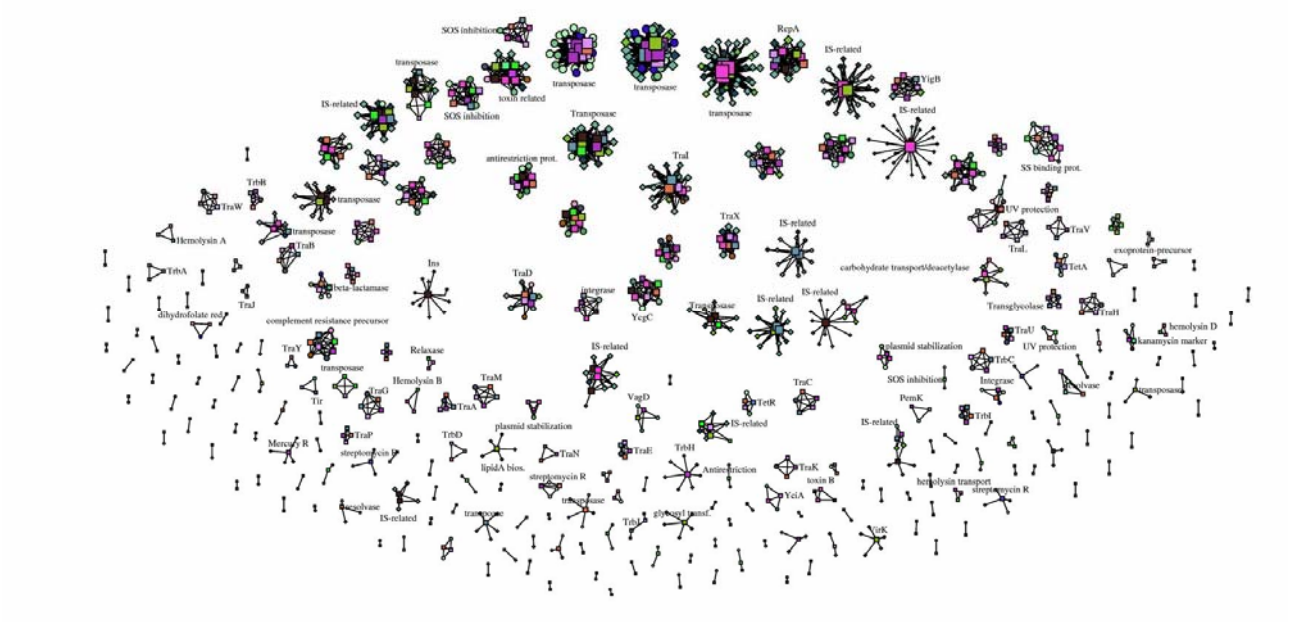
Results

The network representation and analysis of the homologies among plasmid encoded proteins revealed differences when comparing the results obtained for organisms of different taxonomic groups. For example, the average number of links for different plasmid networks is generally lower for free-living organisms, in agreement with an at least partial genetic isolation as the analysis of Sulfolobus plasmids clearly showed, having a little average number of links for node, in agreement with the work of Whitaker et al. (2003). Moreover, the visual inspection of these networks can reveal the presence of large duplicated regions inside plasmids, and the rapid identification of fused proteins. Figure 1 is an example of the uniform network obtained for proteins encoded by Escherichia coli plasmids against a database containing Escherichia coli (squares), Salmonella (circles) and Shigella (diamonds) plasmids encoded proteins (E-value cut-off=0.0001; alignment length cut-off=200; alignment identity cut-off=70% identity); the function, if known, of proteins in the most important clusters is also indicated. In addition of proteins involved in plasmid replication and partition, transposases are the most represented proteins, followed by those involved in conjugative transfer and antibiotic resistanc. Resistance genes are shared by Escherichia coli and Salmonella strains, but not always by Shigella, and this corresponds to the presence of antibiotic resistance containing regions in the chromosome of the last. At this level of filtering, it appears that haemolysin-related genes are peculiar of Escherichia coli plasmids. Moreover, plasmid replication (Rep) proteins appear to be more conserved between Escherichia and Shigella, in agreement with taxonomic relationships.

Contact email: r_fani@dbag.unifi.it

Supplementary informations

After complete testing and refinement the java classes written for this work will be released free for academic users.



Natural cis-antisense of human tumor suppressor genes

Brozzi A (1,2), Pelicci PG (1,2), Luzi L (1,2)

(1) Department of Experimental Oncology, European Institute of Oncology, Via Ripamanti 435, 20141, Milan, Italy.
(2) IFOM, the FIRC Institute for Molecular Oncology Foundation, Via Adamello 16, 20139 Milan, Italy.

Motivation
Large-scale bio-informatic analysis of genomes has recently allowed identification of natural antisense transcripts (NATs) in several model organisms. To elucidate a possible functional significance of cis-NATs in gene expression regulation in cancer we computed the incidence of cis-NAT in a dataset of human tumor suppressor genes and determined their expression in cancer cells by experimental approaches.

Results
At least one antisense transcript was present in 20% of the tumor suppressor dataset. Manual annotation showed that potential coding antisenses are prevalent and that 50% of the cases are conserved in mouse (position and/or sequence conservation). Experimental analyses are on-going in order to determinate the expression profiles of a subset of pairs in some matched normal/tumor samples. Our results show that natural cis-antisense could represent important players in the pathogenesis of cancer through expression regulation on cancer related genes, as tumor suppressors, via double strand editing or chromatin modification.

Contact email: alessandro.brozzi@ifom-ieo-campus.it

Modelling and simulation of the E.coli-Lambda interactions

Cannata N (1), Corradini F (1), Di Berardini MR (1), Mariani F (1), Merelli E (1), Trivelli M (1), Ubaldi M (2)

(1) Dipartimento di Matematica e Informatica, Università di Camerino, Camerino
(2) Dipartimento di Medicina Sperimentale e Sanità Pubblica, Università di Camerino, Camerino

Motivation

Bacteriophage Lambda is a virus that infects E. coli. Upon infection, the phage can propagate choosing between two different programs that depend by environmental conditions. If, for example, the cell is exposed to agents that damage DNA by endangering the existence of bacteriophage itself, the Lambda can activate a program that allows its propagation outside the hosting cell. An approach to understand the behaviour of biological systems consist in the cooperation of biomedical and computer scientists. The integration of the knowledge from the two fields allows the modeling of a biological system of interest and to run the in-silico simulation of this biological prototype in different scenarios. It is the aim of Systems biology [2] to analysis, model and simulate biological systems and processes, through the supply of mathematical and computational models. In the context of LITBIO (http://www.litbio.org) project, we are modelling several biological systems, among those we have analysed and modelled the behaviour of Lambda virus attaching an E. coli bacterium. LITBIO offers a virtual desk, to progressively engineer models of biological systems and to simulate and validate them [3]. The proposed work has been split in two main steps: first, we have studied the literature in order to identify actors, functions and environmental variables involved within the E.coli-Lambda interactions. The second step was dedicated to define the model and simulate the behaviour of the biological system. Thanks to the expressiveness of some informatics tools like modelling languages and timed automata model checking we were able to effectively represents compartments and understand the system dynamics. The final model allows to simulate the activities of Lambda attacking the E. coli and to acquire new knowledge about latency period, infection and virus reaction time.

Methods

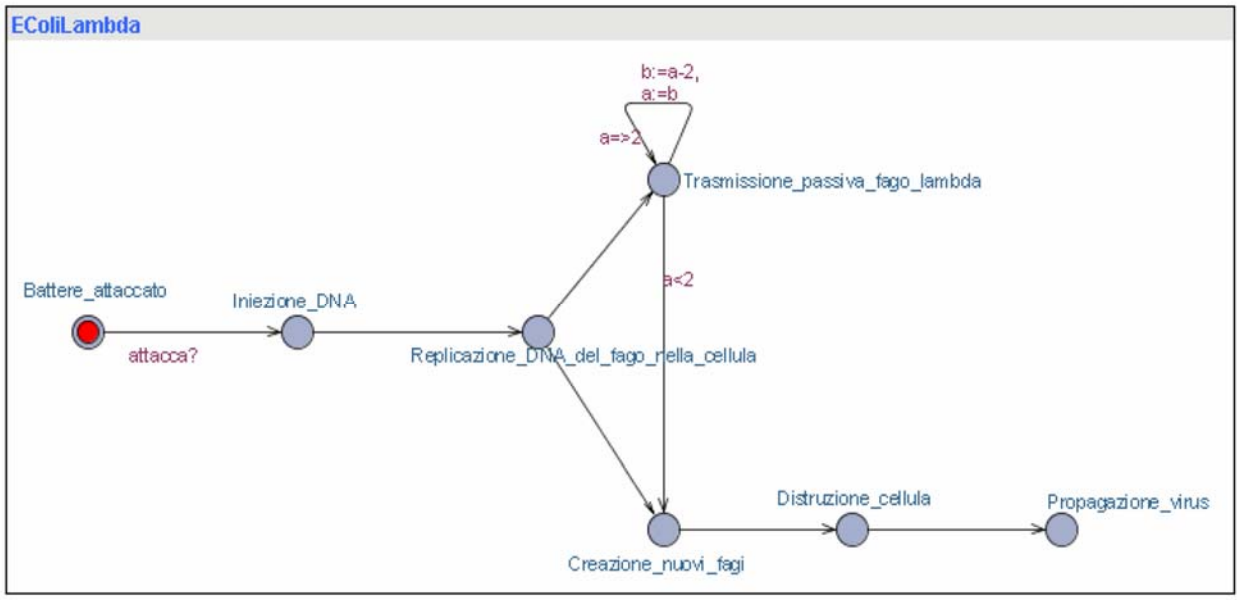
The definition of a model that match as much as possible the reality of a complex biological system requires a step-by-step approach. To this end, computer science offers formal and automatic tools to obtain an intuitive and precise representation of complex systems [4,5]. The E.coli-Lambda system, first has been analysed and represented by UML diagrams. This model allowed a coarse-grain representation of the system through the identification of its main stages. Then, each stage has been analysed with a fine-grain approach. The formal description of the system has been obtained by UPPAL. The UML diagrams help to better understand the interactions that occur between Bacteriophage Lambda and E. coli bacterium, and their autonomous behaviour. The analytical study of the system has been conducted by using the timed automata model checking tool UPPAAL (http://www.uppaal.org). We created a set of finite state automata, to represent the cell compartments autonomous behaviour and the behavoiur of E.coli-Lambda interactions (see Figure). UPPAAL is a graphical tool that provides immediate overview of the dynamic evolution of the system. The use of this tool in the first case of simulation, has been useful to validate the proposed model of the E.coli-Lambda system. Moreover, it provides a more detailed and complete description of the system behaviour.

Results

The UML diagrams and UPPAAL timed automata model checking have been reveled useful tools to obtain an effective model of a biological system of the E.coli-Lambda system. The model allows to study the dynamic behaviour of the systems interactions. The proposed model can be considered the starting point to study and analyse a more complex systems.

Contact email: emanuela.merelli

References
1. Mark Ptashne. A Genetic Switch: Phage Lambda Revisited. Cold Spring Harbor Laboratory Press, 3rd edition, 2004.
2. H. Kitano. Foundations of Systems Biology. MIT Press, 2002.
3. E. Bartocci, N. Cannata, F. Corradini, E. Merelli, L. Milanesi, P. Romano. A multilayer architecture to support bioinformaticians of today and tomorrow, submitted to BITS 2006.
4. N. Cannata, F. Corradini, E. Merelli, A. Omicini, and A. Ricci. An agent-oriented conceptual framework for systems biology. In T. Comp. Sys. Biology, volume 3737 of LNCS, pages 105-122, 2005.
5. F. Corradini, E. Merelli, M. Vita. A multi-agent system for modelling the oxidation of carbohydrate cellular process. In Proc. of the Int. Conference on Computational Science and its Applications, volume 3481 of LNCS, pages 1265-1273, 2005.



Modelling of the eukariotic heat-shock response with probabilistic timed automata

Cannata N (1), Corradini F (1), La Terza A (2), Merelli E (1), Miceli C (2)

(1) Dipartimento di Matematica e Informatica, Università di Camerino, Camerino
(2) Dipartimento di Biologia Molecolare, Cellulare e Animale, Università di Camerino

Motivation

Systems biology [1] aims at the analysis, modeling and simulation of biological systems and processes, through the supply of mathematical and computational models. The heat-shock response (HSR) is an ubiquitous and highly coordinate cellular process which is elicited by eukaryotic as well as prokaryotic cells primarily in response to protein damage [2]. The HSR, including exposure to environmental stresses (xenobiotics, heat shock, heavy metals), pathological states (viral, bacterial and parasitic infections, fever, inflammation, malignancy, ischemia), as well as physiological stimuli (growth factors, tissue development, hormonal stimulation), is characterized at molecular level, by the rapid and transient expression of a specific set of proteins belonging to the evolutionary conserved family of the heat shock proteins (HSPs). The HSPs primarily function as molecular chaperones contributing to protein homeostasis in cells under both normal and stressful conditions. The protective role of these chaperones (HSPs) is mainly exerted at the level of an active participation in the folding, assembly, translocation of proteins (under normal conditions) and in the repair or the degradation of non-native and damaged proteins (under stress conditions). Besides the well characterized roles of HSPs in cell survival and adaptation, to date there are increasing evidences of the importance of their chaperon function in a large numbers of human diseases. Critical roles in the regulation of the HSR are attributed to both the heat-shock transcription factor (HSF) and one of the major HSPs, the heat-shock protein 70 (Hsp70). Under normal conditions, the HSF is bound to Hsp70 in the cytoplasm of mammalian cells. Under stress conditions such as heat-shock and ischemia, HSF is separated from Hsp70, rapidly converted from a monomer to a trimer and inducibly phosphorylated and concentrated in the nucleus to activate heat-shock gene transcription. The newly synthesized HSPs bind to HSFs to prevent further synthesis of HSPs via an auto-regulatory loop. The longterm aim of this study is to clarify the molecular mechanism(s) by which the eukaryotic cells sense and respond to stress. Although the same set of genes, known as heat shock genes are induced, different members of the HSF family are activated. All of them bind to conserved genetic elements known as heat-shock elements (HSE). We are currently developing a predictive model of the cytoplasmic and nuclear events of the eukaryotic HSR by means of probabilistic timed automata [3] trying to gain a better understanding of this universal cellular phenomenon.

Methods

We propose to model the structure and the dynamics of heat shock response of a cell as a probabilistic timed automata, a formal description mechanism in which both nondeterministic and probabilistic choices can be represented. In our model the compartments (cytoplasm and nucleous) of the cell are automata, and the actions are the interactions among components performing cellular processes. The initial stages of the modelling process employ the timed automata model checking tool UPPAAL [4] to automatically verify the soundness of several abstractions applied to our probabilistic timed automaton model. The use of nondeterminism allows us to model asynchronous behaviour of cell components. Moreover, the use of probabilities give us a chance to verify properties referring both to the likelihood of interaction with proteins, and to the probability that a cell component reacts to a certain environmental event.

Results

The model of eukariotic heat shock response has been created with probabilistic timed automata which allowed to clarify the molecular mechanism(s) by which eukaryotic cells sense and respond to stress. It is expected that the same set of genes (heat shock genes) are induced by different

members of the HSF family; HSF1 becomes active in response to classical stress stimuli, whereas HSF2 is active during differentiation-related processes.

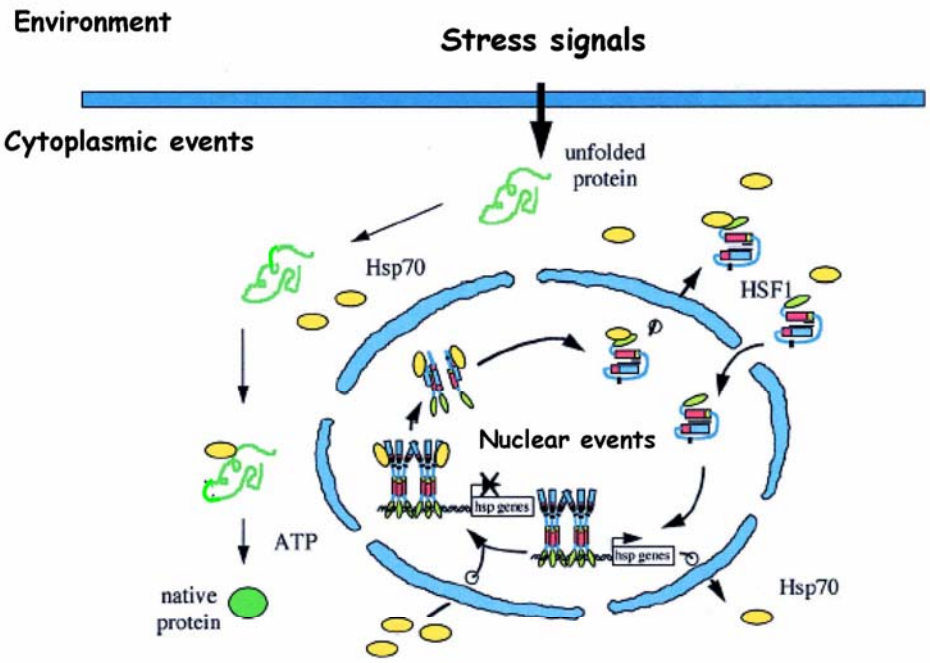
Contact email: emanuela.merelli@unicam.it

References

1. H. Kitano. Foundations of Systems Biology. MIT Press, 2002.
2. M.G. Santoro. Heat shock factors and the control of the stress responce. Biochemical Pharmacology, vol. 59, pp.55-63, 2000.
3. M. Kwiatkowska, G. Norman, R. Segala, and J. Sproston. Automatic verification of real-time systems with discrete probability distributions. Theoretical Computer Science, 286, 2002.
4. K. G. Larsen, P. Pettersson, and W. Yi. UPPAAL in a nutshell. Software Tools for Technology Transfer, 1(1+2):134-152, 1997.
5. E. Bartocci, N. Cannata, F. Corradini, E. Merelli, L. Milanesi, P. Romano.

Supplementary informations

A multilayer architecture to support bioinformaticians of today and tomorrow, submitted to BITS 2006. The proposed study is being carried within the MIUR-FIRB LITBIO project (http://www.litbio.org/).



Regulation of the heat shock response (modified from Santoro MG, Biochem. Pharmacol., Vol.59, pp. 55-63, 2000

A resource ontology for bioinformatics Resourceomes

Cannata N, Corradini F, Leoni L, Merelli E, Piersigilli F

Dipartimento di Matematica e Informatica, Università di Camerino

Motivation

Bioinformatics constitutes a very dynamic and multidisciplinary galaxy in the scientific universe. At the dawn of the "omics" age it was defined as a discipline intended to manage and analyze the exponentially growing amount of biological data. Nowadays it has become hard to image genomics, systems biology, and other emerging fields of the modern biology, disjointed from bioinformatics. Bioinformatics itself represents one of the forces driving the paradigm shift of science toward e-Science. Beside "data deluge" and "information overflow", it's straightforward to figure out a similar condition also for knowledge. Actually we are witnessing a fast increasing production of scientific literature and the blooming of knowledge representation efforts (i.e. biomedical ontologies). Therefore, we could metaphorically depict bioinformaticians drowning in the ocean of resources (e.g. databases, articles, programs) developed from themselves. In general, the quest for the "right" resource has become a very demanding and time-consuming task. It's almost impossible for a human scientist to follow the evolution of the general field. The tracking of new resources, appeared in a limited research sector, represents already a challenging issue. In [1] we defined "resourceome" the full set of bioinformatics resources and invited the community to organize a general, machine-understandable, index of bioinformatics resources. Such an index should also take into account the semantic relationships between resources.

Methods

Here we adopt the same term but with the uppercase R Resourceome, to identify the ontological representation of a typically huge set of resources. Aiming at the fulfillment of a general index of bioinformatics resources, we can foresee its bottom-up formation process. Individuals and special interest groups build their limited and probably overlapping Resourceomes, which subsequently could be shared, merged and integrated into a general one. In a Resourceome the knowledge of a domain is organized into an ontology, and the resources related to a concept of the domain are directly connected to that concept. "In-vivo" and "in-silico" scientists can easily navigate and "reason" through the semantic networks connecting resources and domain's concept. A single scientist could adopt a Resourceome to intuitively organize his/her perceived knowledge of a domain and the related preferred resources. We can naturally image Resourceomes also at the heart of next generation bioinformatics cyberinfrastructures. The adoption of semantic web technologies permit to arrange the concepts of the domain and resource's metadata, together with their relationships. Software agents, beside supporting the building and management of the ontology, permit to keep it "alive" and will support users in the navigation and queries.

Results

We can reasonably assume that most of the peer-reviewed articles published in bioinformatics journals are intended to present new databases, algorithms, programs and other computational resources. Therefore, to provide also numerically the perception of the growth of the bioinformatics resourceome in the last years, we performed an analysis on the number of new bioinformatics journal and on the amount of published pages and articles for the most significant ones. Then, we developed a resource ontology to classify bioinformatics resources and to exploit their typical semantic relationships. Such a general classification, orthogonal to the domain ontology, is intended to be adopted in Resourceomes to describe, also visually, the type of the resources connected to concepts of the domain.

Contact email: nicola.cannata@unicam.it

References

1. N. Cannata, E. Merelli, and R. B. Altman. Time to organize the bioinformatics resourceome. PLoS Comput Biol., 1(7):e76, 2005.

Supplementary informations

This work is supported by the FIRB project Laboratory of Interdisciplinary Technologies in Bioinformatics (LITBIO). We are grateful to Russ B. Altman for strongly supporting the Resourceome idea.

MS-Analyzer: Composing and Executing Preprocessing and Data Mining Services for Proteomics Applications

Cannataro M, Veltri P

University Magna Graecia of Catanzaro, Catanzaro, Italy

Motivation

Mass Spectrometry (MS) proteomics produces huge datasets, said spectra, that contain large set of measures (intensity, m/Z), representing the abundance of biomolecules having certain mass to charge ratios. MS data hides a lot of information about cell functions and disease conditions and can be used for various analysis, e.g. biomarker discovery, peptide/protein identification, and sample classification. The discovering of such information needs the combined use of bioinformatics and data mining, and requires the efficient access to huge spectra datasets and various software tools for to the loading, management, preprocessing, and mining of spectra, as well as the interpretation and visualization of discovered knowledge models. The increasing use of MS in clinical studies causes the collection of spectra data from large sample populations, e.g. to control the progression of a disease. In addition, the comparative study of a disease may require the analysis of spectra produced in different laboratories, so it is possible to envision that in few years biomedical researchers will need to collect and analyze more and more spectra data. Since spectra have a high dimensionality and are often affected by errors and noise, specialized spectra databases and preprocessing techniques are needed. Finally, MS involves different technological platforms, such as sample treatments, MS techniques, spectra processing, data mining analysis, and results visualization. Choosing the right methods and tools requires multidisciplinary knowledge from MS specialists to biologists and computer scientists, thus, modelling the semantic of processes, tools, and data is a key issue to simplify application design.

Methods

Ontologies constitute a well established tool to model the steps of data mining applications and support the application design, while Grid technology may provide: efficient storage space where maintaining on line large spectra datasets, broadband infrastructure needed to collect in a secure and efficient way proteomics data coming from remote laboratories, computational power needed by preprocessing and mining algorithms. To address the key issues of spectra data management and analysis we provide the basic MS bioinformatics tools as Web Services and propose the combined use of ontologies for the modelling of protemics tools, workflow techniques to compose in a seamless way basic proteomics services, and the Grid as the deployment infrastructure. The proposed bioinformatics platform, named MS-Analyzer, offers to the biologist a set of high level services, namely: spectra management services, providing spectra format conversion and efficient spectra storage through a specialized spectra database; moreover experimental data are modelled through the dataset concept, i.e. a set of spectra that can be in raw, preprocessed or prepared stage; pre-processing services, that implement common spectra pre-processing algorithms, such as base line subtraction, smoothing, normalization, binning, peaks extraction, and peaks alignment; data preparation services, that provide the spectra reorganization needed when applying data mining tools (e.g. Weka tools require spectra dataset formatted in a unique input file having a specific metadata header); data mining services, obtained by wrapping Weka tools, a popular data mining suite; moreover, tools for knowledge models visualization are provided. An Ontology-based Workflow Editor allows the concept-based browsing/searching of such services modelled through the MS-Analyzer ontologies: WekaOntology, that models the Weka data mining tools and is enriched by the description of relevant spectra concepts and pre-processing algorithms, and ProtOntology, that models concepts, methods, algorithms, tools, and databases relevant to the proteomics domain, and provides the biological background and perspective to the data mining analysis.

Results

The paper presented MS-Analyzer, a Grid-based software platform that supports the semantic composition of spectra preprocessing algorithms, efficient spectra management techniques based on a specialized spectra database, and off-the-shelf data mining services, to analyze mass spectra data on the Grid. By using MS-Analyzer a user can easily design a data mining application with the help and the constraint checks provided by WekaOntology and ProtOntology, and without worrying of software details, having a suite of specialized spectra management services that simplify and automate the path to knowledge discovery. The availability of different services allows to produce in few time many workflows of the same application employing different combination of preprocessing, preparation and data mining techniques. The produced knowledge models as well as the execution performance of the scheduled workflows can be easily visualized, allowing to compare the effect of different preprocessing and data mining techniques and to evaluate the best strategies to analyze mass spectra data.

Contact email: cannataro@unicz.it

Entropic Embedding in High-Dimensional Biological Systems

Capobianco E

Department of Biomedical Engineering, Boston University, Boston

Motivation
The exploratory analysis and computational modeling of complex high-dimensional systems represent important interdisciplinary research areas for many application fields. Challenging inference problems that are often addressed concern model building and variable selection, clustering and feature extraction, non-linear structure detection and relationship between observed and intrinsic dimensionality, signal extraction, de-noising, and network dynamics reconstruction. The attempt of this work is to leverage on such methodological wealth for dealing with problems relevant to systems and computational biology.

Methods
Genomic data are examples of noisy high-dimensional systems whose observed dynamics may be viewed as mixtures of informative sources with unknown statistical distribution and subject to unknown mixing mechanism. Gene expression values have many interesting features that depend on complex network dynamics. This work presents an application of independent component analysis (ICA) interestingly combined with fuzzy rules, embedding principles and entropic measures.

Results
Entropy and embedding turn out to be very useful tools for controlling the robustness and stability of the decomposition of a system with larger than intrinsic dimensionality in the observed variables, and complex dynamics in the hidden ones. We report results from experiments that show convergence to the intrinsic dimensionality from the observed genomic space by the means of its least dependent decomposition in just a minimal number of salient features.

Contact email: enrico_capobianco@yahoo.com

Supplementary informations
The author is currently on leave and re-locating to a new research Institute. This work was conducted completely at Boston University.

A systematic approach for noise characterization in ESI-Q-TOF spectra

Cappadona S (1), Pattini L (1), Levander F (2), James P (2), Cerutti S (1)

(1) Dipartimento di Bioingegneria, Politecnico di Milano, Milano, Italy
(2) Department of Protein Technology, Lund University, Lund, Sweden

Motivation

HPLC-MS/MS is increasingly becoming the method of election for large-scale proteomic analysis. The sensitivity of protein identification by peptide sequencing is strongly dependent on the ability to discriminate low intensity peptide signals from the underlying noise. A filtering step preceding peak detection may improve the identification results, but needs a general investigation into the nature of noise and how it can be reduced or filtered. The aim of this work is to develop a systematic approach to characterise and eliminate noise, rather than characterise and detect peaks, as principally proposed in literature.

Methods

A four step approach has been specifically formulated in order to characterise the noise. 1. Acquisition, by mean of an ESI-Q-TOF mass spectrometer, of a dataset of spectra from samples produced ad hoc. This dataset includes a blank run, without peptides, and a set of runs in which only few and known peptides have been placed in the samples. 2. Conversion of the spectra from their original m/z domain back to the time domain (that of the times of flight of ions in the spectrometer). The first domain is ideal for the characterization of the chemical noise, while the second one is more appropriate for the periodic one. 3. Fourier transformation of the m/z spectra in order to find common components of the chemical noise all over the spectra. 4. Fourier transformation of the time spectra in order to find common components of the periodic noise all over the spectra.

Results

The analysis of the spectra of the blank run after Fourier transformation has revealed the existence of components which are present all over the run. Since the blank doesn't contain peptides, these components are clearly typical noise features. Moreover, the same components have also been detected in all of the spectra from the other runs which do not contain peptide peaks. These results suggest that our approach can be used to characterise typical components of the chemical and the periodic noise. The characterization of the features of the noise can be used as the first step for noise detection and rejection.

Contact email: salvatore.cappadona@biomed.polimi.it

ESTuber db: a tool for Tuber borchii EST sequence management

Caprera A (1,5), Cosentino C (2,6), Viotti A (2), Stella A (3), Milanesi L (4), Lazzari B (2,5)

(1) CISI, Via Fratelli Cervi 93, 20090 Segrate (MI), Italy
(2) Istituto di Biologia e Biotecnologia Agraria, via Bassini 15, 20133 Milan, Italy
(3) Parco Tecnologico Padano, Via Einstein - Località Cascina Codazza, 26900 Lodi, Italy
(4) Istituto Tecnologie Biomediche, Via Fratelli Cervi 93, 20090 Segrate (MI), Italy
(5) Current address: Parco Tecnologico Padano, Via Einstein - Località Cascina Codazza, 26900 Lodi, Italy
(6) Current address: Darmstadt University of Technology, Institute of Botany, Schnittspahnstrasse 3-5, 64287 Darmstadt, Germany

Motivation
The ESTuber database (http://www.itb.cnr.it/estuber/) represents a collection of 3,271 expressed sequenced tags (EST) of the white truffle Tuber borchii. The dataset consists of 2,389 sequences from an in-house prepared cDNA library obtained from vegetative mycelium, and 882 sequences downloaded from GenBank, representing four libraries obtained from vegetative hyphae and fruit bodies at different developmental stages. An automated pipeline was prepared to process EST sequences by using public software integrated with in-house developed Perl scripts. Data produced during EST processing were parsed and collected in a MySQL database. The database can be queried via a php-based web interface. The aim of this work was to create a public comprehensive resource of data and links related to truffle EST sequences.

Methods
A multifasta file containing the complete truffle EST dataset was used as input for the CAP3 program and 356 contigs were generated. Extensive sequence annotation was performed by blastx against the GenBank nr protein db and by blastn against an in-house prepared database of more than 42,000 genomic sequences from four filamentous fungi (Magnaporthe ssp, Aspergillus ssp, Fusarium ssp and Neurospora ssp) and a dimorphic fungus (Saccharomyces ssp). Blastx was also performed against the UniProtKB database, to annotate sequences according to the Gene Ontology (GO) project, and an algorithm was implemented to infer statistical classification from the ontologies occurrences. GO statistics are provided on the ESTuber db web site for the whole sequence set and for library-specific subsets. All the blast output pages are available and can be queried by text search. Detection of tandem repeats was performed on all the EST sequences, as well as on the contig consensus sequences, with the Tandem Repeats Finder software. Putative polypeptide sequences were deduced from nucleotide sequences with FrameFinder and compared to the PROSITE database of protein families and domains. Links to matching patterns are given in the database web interface. A local blast utility was set up to perform blast searches on the ESTuber db nucleotide dataset or on the inferred protein sequences. A text search utility allows querying all the database fields and query outputs can be downloaded in multiple formats. Logical sequence subsets (singlets/sequences participating to contig assembly, unigenes, repeats containing sequences, protein pattern containing sequences) were defined and can be searched as independent datasets. An exhaustive help page was included in the database to help users to surf the database and to interpret the program outputs.

Results
The resulting database represents, at present, the most comprehensive public resource for Tuber EST sequences, and can be helpful to all researchers interested in the molecular biology of truffle and of other filamentous fungi. The Perl pipeline automatically fills in all the database fields and very little manual participation is required for the complete assembly of new releases. The database structure is modular and new applications can easily be integrated. The in-house developed GO statistics tool was useful for comparison of inter-species ontologies occurrances as well as for database data analysis, and represents a powerful tool for investigating EST distribution in Gene Ontology categories. The ESTuber db is intended to be the main repository of information related to Tuber sequences, and will be updated and modified according to the needs of the Truffle Project.

Availability: http://www.itb.cnr.it/estuber/

Contact email: barbara.lazzari@tecnoparco.org

Improving the quality of the predictions of protein stability changes upon mutation using a multi-class predictor

Capriotti E (1), Fariselli P (1), Rossi I (1,2), Casadio R (1)

(1) Laboratory of Biocomputing, CIRB/Department of Biology, University of Bologna, via Irnerio 42, 40126 Bologna, Italy
(2) BioDec Srl, Via Calzavecchio 20/2, 40033 Casalecchio di Reno (BO), Italy

Motivation

The accurate prediction of protein stability free energy change (DDG) upon single point mutation is a key problem of Structural Bioinformatics. In the last years several methods were described to predict DDG upon single point mutations in proteins. One common approach is based on the development of different energy functions, starting routinely from the protein known structure and then applied to the mutated protein to compute DDG. Recently the increasing number of experimental thermodynamic data and their availability in the ProTherm database prompted us to develop machine learning-based approaches for predicting both the sign and the value of DDG upon protein mutation starting both from the sequence and/or structure (I-Mutant2.0, http://gpcr2.biocomp.unibo.it/cgi/predictors/I-Mutant2.0/I-Mutant2.0.cgi).
These automatic methods however suffer from the fact that experimental data are affected by standard deviations associated with the DDG values, when evaluated, and from the fact the most of the experimental data (about 32% of the data set) are close to 0 (-0.5 =< DDG =< 0.5 Kcal/mol). In these cases, considering the associate error, both the value and the sign of DDG may be either positive or negative for the same mutation, leading to ambiguity when evaluating the extent of protein folding stability. In order to overcame this problem we implemented a new predictor able to discriminate between 3 possible classes, dividing the set of experimental data in: destabilizing mutations, stabilizing mutations and neutral mutations. Furthermore we also enriched the training set of experimental data by assuming that for each mutation in the data base also the opposite restoring mutation is present.

Methods

The databases used in this work are derived from the release (September 2005) of the Thermodynamic Database for Proteins and Mutants ProTherm. We select our initial set imposing the following constrains: a) the DDG value was extrapolated from experimental data and reported in the data base; b) the data are relative to single mutations; c) the data are obtained from reversible experiments After this procedure we obtain a larger data set comprising 1681 different single point mutations and related experimental data for 58 different proteins. From the latter by selecting only 55 protein known with atomic resolution we have a subset of 1634 mutations. Adopting a criterion of thermodynamic reversibility for each mutation, we double all the thermodynamic data. Finally, we end up with 3362 mutations for the set containing protein sequences (DBSEQ) and 3268 mutations for the subset of proteins known with atomic resolution (DB3D). According to experimental DDG value each mutation is grouped into one of the following three classes: i) destabilizing mutation, when DDG<-0.5 Kcal/mol; ii) stabilizing mutation when DDG>0.5 Kcal/mol; iii) neutral mutations when -0.5 =< DDG =< 0.5 Kcal/mol. The choice of |0.5| Kcal/mol as a threshold value for DDG classification provides a balanced datasets and is also a limiting value of standard errors reported in experimental works. We developed support vector machines (SVM) and trained them to predict if a given single point protein mutation is classified in one of the three classes defined above. This task is addressed starting from the protein tertiary structure or sequence, adopting a Radial Basis Functions kernel. The input vector consists of 42 values. The first 2 input values account respectively for the temperature and the pH at which the stability of the mutated protein was determined. The next 20 values (for 20 residue types) explicitly define the mutation (we set to -1 the element corresponding to the deleted residue and to 1 the new residue (all the remaining elements are kept equal to 0). Finally, the last 20 input values encode the residue environment: namely a spatial environment, when the protein structure is available, or the nearest

sequence neighbours, when only the protein sequence is provided. When prediction is structure-based, the Relative Solvent Accessible Area (RSA) is also coded as input value.

Results

Our methods are trained/tested using a 20-fold cross-validation procedure on the two available sets. After an optimization procedure on different spatial and sequence environments, best predictors score as follows: when based on structural information, an overall accuracy of 58% is achieved with a mean value of correlation to the thermodynamic data of 0.37. In turn, when the prediction is performed considering only sequence information the accuracy is 52% and the mean value of correlation becomes 0.28.

Availability: http://www.biocomp.unibo.it/

Contact email: emidio@biocomp.unibo.it

A new interactive and fully automated tool for parsing BLAST output

Carreras M (1), Bosotti R (2)

(1) Bioinformatic Software Developer, Pavia
(2) Genomics Unit, Department of Biotechnology, Nerviano Medical Sciences, Nerviano (MI)

Motivation
Blast is worldwide acknowledged as a strong research instrument and nowadays it has evolved to generate new Blast types such GEO and SNP Blast. For large datasets, manual inspection of Blast output is inadequate and parsing tools are needed to see valuable information in a more compact structure.

Results
We present a new BLAST parser tool providing a user-friendly interface to manage the output of all these newer BLAST types. The output is organized in a grid structure, that can be easily exported to several formats, recognized by most the existing applications and database managers.

Availability: http://geneproject.altervista.org/

Contact email: mailto:geneproject@altervista.org

Improving the selection of close-native protein structures in decoy sets using a graph theory-based approach

Casadio R (1), Fariselli P (1), Margara L (2), Filippo M (2), Vassura M (2)

(1) Biocomputing Group Department of Biology , University of Bologna.
(2) Computer Science Department, University of Bologna.

Motivation

One of the still unsolved problems in the ab initio protein structure prediction is the ability of distinguishing from near-native and distant-native protein structures. Indeed ab initio methods generate several structures, often spanning all the known structural types and in the absence of the known real solution (the real protein structure) it is very difficult to select the most likely fold/s of a given chain. Due to the intrinsic errors of the force fields presently available, several filtering procedure have been therefore developed and implemented, including combinations of different energy functions (1). The problem at hand can be addressed only when for a given protein, a good decoy set is also available (1). This is necessary in order to test the discriminative ability of the different methods

Methods

Here we take a new approach introducing a graph representation of each protein and associated decoys. The data set of selected proteins and decoys is somewhat modified from that computed in a previous published work (1). In (1) good decoys were computed and made available following a very stringent definition of "good" decoy set. Decoys were produced with the Rosetta method for a large set of proteins, following four criteria: 1) contain conformations for a wide variety of different proteins; 2) contain conformations close (<0.4 nm) to the native structures; 3) consist of conformations that are at least near local minima of a reasonable scoring function; 4) be produced by a relatively unbiased procedure. We consider a set of 41 proteins with 1800 decoys per protein, for a total of about 76000 protein structures. Given the data set of protein and decoy structures, we consider fifteen properties taken from graph theory and we test their ability to distinguish correct from incorrect folds. Our method stems from the notion that according to each property decoys can be ranked with respect to the corresponding protein structure as a function of structural similarity. In our approach each protein and related decoys are represented with a graph adjacent matrix or "contact map". For each decoy set (or set of decoy contact maps per each protein) we computed the number of edges (number of contacts), average degree (average number of contacts per residue), contact order, diameter, complexity, flow, connectivity, and also the variances and weighted versions of each property, respectively. For each property and for a given decoy set, we evaluated the Enrichment measure introduced by Tsai et al. (1) and the Z score. The ability of a given graph property to act as a scoring function is then evaluated by computing the Enrichment measure and Z score (1). More specifically, to compute the Enrichment due to a given property, we count how many of the best decoys are among the best decoys according to Ca-RMSD (Root Mean Square Deviation of the C alpha backbone of the decoy to that of the corresponding protein) and compare this number with what would be expected for a uniform distribution. Due to redundancy of decoy structures, it may occur that very similar structures are present in different decoy sets. To avoid this overlapping in the final evaluation, we keep into consideration only those decoys found in the intersection set between the top high scoring 15% decoys, as obtained according to the property at hand, and the top 15% decoys with the lowest Ca-RMSD for a given protein. This figure is then divided by the number of random assignment (15%*15%. number in the set) to highlight the performance of the method. Values greater than one indicate an Enrichment over a uniform distribution.

Results

We evaluate the Enrichment and Z-score due to each selected graph functions and we obtain that according to some of these properties (such as connectivity, network flow, and particularly

complexity) the selection of optimal decoys outperforms previously described methods based on a combination of all-atom energy functions (1). With complexity the score is 1.4 times better than with other methods. We suggest that our approach can be applied when in "ab initio" predictions a set of "good structures" need to be selected among a population of predicted structures. This method can therefore complement the classical approach of energy-based scoring functions and help in solving the protein folding problem.

Contact email: casadio@alma.unibo.it

References

1. J Tsai, R Bonneau, AV Morozov, B Kuhlman, CA Rohl, and D Baker, "An improved protein decoy set for testing energy functions for protein structure prediction." Prot Struct Func Genetics, 2003

The MEPS server for identifying protein conformational epitopes

Castrignanò T (1), D'Onorio De Meo P (1), Carrabino D (1), Orsini M (2),
Floris M (2), Tramontano A (3,4)

(1) CASPUR, Consorzio Interuniversitario per le Applicazioni di Supercalcolo per Universita` e Ricerca, Roma
(2) Center for Advanced Studies, Research and Development in Sardinia (CRS4), Bioinformatics Unit, PULA (CA)
(3) Department of Biochemical Sciences, University 'La Sapienza', Roma
(4) Istituto Pasteur-Fondazione Cenci Bolognetti, University 'La Sapienza', Roma

Motivation
One of the most interesting problems in molecular immunology is epitope mapping, that is the identification of the regions of interaction between an antigen and an antibody. The solution to this problem, even if approximate, would help in designing experiments to precisely map the residues involved in the interface and could be instrumental both in designing peptides able to mimic the interacting surface of the antigen and in understanding where important regions of an antigen are located in its three-dimensional structure.

Methods
We have developed and implemented a method that, given the structure (or a model of the structure) of an antigen or of a set oif antigens identifies all peptide sequences able to mimic the surface of the antigen.

Results
The server can either provide an exhaustive list of all peptides of a given length, or search for possible sequence on the surface of the protein, given the amino acid sequence of a fragment and a maximum number of allowed mismatches. In the first case, sequences corresponding to putatively mimicking peptides are stored, together with information about their location in the protein sequence and structure, as a FASTA formatted file. This allows users to directly search the seqeunecs with the many widely available programs for data base searching. BLAST searches are directly available via the server.

Availability: http://www.caspur.it/meps

Contact email: anna.tramontano@uniroma1.it

Design of Selective Compounds for the Estrogen Receptors: A Molecular Docking and Machine Learning Study

Chiappori F, Ferrario MG, Gaiji N, Fantucci P

Department of Biotecnology and Bioscience, Università Milano - Bicocca, Milano DELOS s.r.l, Bresso (MI)

Motivation

Estrogens are steroidal hormones that regulate reproductive functions and the bone mass maintenance in mammals. The molecular action of the estrogens is mediated by the nuclear receptors hERs, isoforms ERa and ERb, which show high structural similarity, but low sequence identity. In women, after menopause the production of estrogens is strongly reduced, it is partly recovered by the Hormone Replacement Therapy (HRT) that prevents some of the menopause symptoms, but enhances the risk of tumors of the reproductive tissues. The most promising model compounds for HRT are SERMs (Selective Estrogen Receptor Modulators) and phytoestrogens. SERMs act in a tissue-specific manner, but they are not sufficiently strong agonists in target tissues, while in other tissues they are antagonists; phytoestrogens are not tissue-selective, but they act mainly on the ERb, which seems to be responsible for the non-reproductive estrogen functions. Our research aims at identifying new compounds for HRT, using a Virtual High-Throughput Screening (VHTS) approach, based on generation of a large number of drug-like compounds and analysis of their interaction with the Ligand Binding Site (LBS) of the ERa and ERb. The 3D structures of the two proteins have been obtained as described in another contribution to BITS (1). All the results reported have been obtained using the DELOS suite of programs (2) which has been recently developed by us. The LBSs found by DELOS correspond very closely to the binding cavity of the ERa and ERb, experimentally known (3).

Results

Two compounds libraries have been generated using as scaffolds the skeletons of estradiol and genestein, respectively, on which three substitution points have been defined for ten different substituents. The two libraries, named EST and GEN, respectively, include 1000 compounds each. The DELOS program has been used for optimization of the 3D structures of the ligands using a quantum chemical approach and for the evaluation of 346 molecular descriptors for each compound. Another section of DELOS allows the construction of the potential maps for the ligand-LBS interaction, to be used in rigid docking simulations. The rigid docking module of DELOS is based on a fast Genetic Algorithm (GA) coupled with an extrapolation scheme (GA-E) for the global optimization of the best roto-translation of the ligand within the LBS. On the basis of the differential docking energy to ERa and ERb (minimum with ERa and maximum with ERb) 65 and 57 molecules have been selected and considered as "good molecules" from EST and GEN libraries, respectively. All other elements of the libraries have been considered as "bad molecules". This coarse classification is carried out in order to attempt a two-way classification of the library elements, using the molecular descriptors as features and the docking behaviour as class variable. We submitted to the classifier the table of the all 346 descriptors, after that, we tried different feature selection methods: · the descriptors with non-zero variance (264 for EST and 290 for GEN)
· the descriptors with correlation index above 0.4 for the docking energy to ERα (68 for EST and 111 for GEN) and to ERβ (84 for EST and 94 for GEN), and for the difference of docking energy from ERα to ERβ (12 for EST and 40 for GEN). The results of the classifier trained by the EST library are good in the case of the descriptors with non-zero variance (74% true positives, 77% true negatives) and the descriptors correlated with the docking energy for ERβ (80% true positives, 61% true negatives). In the case of GEN library, the best results are those provided by the classifier trained by the descriptors with non-zero variance (90% true positives, 65% true negative) and the descriptors correlated to the docking energy for ERβ (88% true positives, 52% true negatives). The final selection performed on GEN and EST libraries was carried out using a combined criterion which incorporate both docking energy and a series of ADME properties. The final subset of

molecules includes 12 and 8 molecules for EST and GEN library, respectively. Concerning the research of molecules for the HRT, we have individuated some characteristics that may play a major role for finding new ERβ-selective compounds, i.e. the absence of bulk side chains, the presence of bulk substituents with polar groups at the ends. The present investigation shows how our VHTS procedure is efficient and relatively fast. The bioinformatics platform DELOS, which allows to follow all the important steps along the rational drug design process, proved to be a reliable and very promising tool for massive, extended studies in this field.

Contact email: noura.gaiji@unimib.it

References

1. Molecular dynamic study of the ligand binding domain of estrogen receptor alpha and beta. Chiappori F., Fantucci P., Ferrario MG, Gaiji N BITS2005
2. DELOS is an integrated platform H/S developed by the academic spin-off DELOS.srl
3. Brzozowski et all Nature vol 389, 753-758 (1997)

A systems biology approach to the functional screening of genomes

Chiarugi D (1), Degano P (2), Marangoni R (2)

(1) Department of Mathematics and Informatics, University of Siena, Siena.
(2) Department of Informatics, University of Pisa, Pisa.

Motivation

Comparative genomics usually manages the functional aspects of genomes, simply by the comparison of gene-by-gene functions. Following this point of view Mushegian and Koonin (Mushegian A.R. and Koonin E.V. "A minimal gene set for cellular life derived by comparison of complete bacterial genomes." Proceedings of National Academy of Science USA, 93:10268-10273, 1996) proposed a hypothetical minimal genome, obtained by eliminating duplicate or apparently functionally identical genes from the genomes of very simple contemporary bacteria, with the aim to find a possible very ancestor genome. The Authors were unable to answer the fundamental question: is, such a hypothetical organism, able to live? We performed a dynamic simulation of the metabolic activities of this hypothetical organism in order to assess whether it is at least able to live virtually, or, in other words, if the dynamic simulation of a virtual cell with this genome will lead to some equilibrium state.

Methods

We specified all the enzymes encoded by the proposed gene set and wrote down all the metabolic reactions which could take place in that hypothetical cell. We represented all metabolites, enzymes and reactions using the enhanced pi-calculus formal language. We developed a simulator written in JAVA, and we used this simulator to test whether the formalized cell is able to reach an equilibrium state or not. Further details on the formalization method and procedure can be found in: Chiarugi D., Curti M., Degano P., Marangoni R., "ViCe: a Virtual Cell", Lect. Notes Comp. Sc., 3082: 207-220, 2005.

Contact email: marangon@di.unipi.it

Supplementary informations

Our simulation clearly shows that the minimal gene set (MGS) genome proposed by Mushegian and Koonin does not express an organism able to live. We proceeded by progressive functional replacements in order to find a genome composition able to give rise to an equilibrium. Among the original 256 genes in MGS, 75 of them have been ruled out because they are functionally duplicated. For example, some MGS genes encoding for proteins involved in the uptake of extracellular metabolites, as in the case of the uptake and phosphorilation of glucose. In this case we choose to maintain only the PTS system, because it is characteristic of Bacteria. Moreover, we needed to add 6 genes not present in MGS and coding for enzymes involved in critical nodes of the metabolic network, as in the case of ribulose-5-phosphate isomerase which leads to the synthesis of D-ribose-5-phosphate, a not dispensable metabolite for nucleotide synthesis. After these modifications, the 187-genes resulting genome is able to give rise to a virtual living organism. Moreover, the distribution of the concentrations of virtual metabolites, once the equilibrium is reached, is very similar to that experimentally measured in bacteria.

An integrated platform for solanaceae genomics

Chiusano ML (1), D'Agostino N (1), Traini A (1), Raimondo E (1),
Aversano M (1), Frusciante L (2)

(1) Department of Structural and Functional Biology, University 'Federico II', 80134 Naples, Italy
(2) Department of Soil, Plant and Environmental Sciences, University 'Federico II', 80055 Portici (NA), Italy

Motivation
We present here our effort as partners of the Solanaceae (SOL) Genomics Network. The long term goal of the Consortium is to build a network of resources and information dedicated to the biology of the Solanaceae family which includes many species of relevant agricultural interest such as tomato and potato. In the frame of the International Tomato Sequencing Project (http://www.sgn.cornell.edu/help/about/tomato_sequencing.pl), to efficiently exploit genomics data generated within the Consortium, the Bioinformatics Committee coordinates data management and integration and offers analysis tools in a distributed platform to support the research of all the SOL partners. As participants to the Bioinformatics of the Network, in the first year project we set up a workbench to support the experimental annotation of the Solanum lycopersicum (tomato) genome and the analysis of sequence collections from other Solanaceae. The workbench under development as well as some relevant issues related to a large scale genomic effort are presented.

Methods
Using an automated approach (D'Agostino et al, BMC Bioinformatics 2005) we built and maintain two databases of ESTs from tomato and potato species (D'Agostino et al., BITS 2006) downloaded from dbEST. The databases provide a complete set of computationally defined transcript indices representing unique sequences, singletons or tentative consensus (TC), derived from EST clustering analysis. We based the annotation of the expressed sequences on the use of controlled vocabularies such as the Gene Ontologies (GO; The Gene Ontology Consortium 2000) and the Enzyme Commission numbers (EC; Bairoch, 2000) and implemented the 'on the fly' mapping of the expressed sequences onto known metabolic pathways from KEGG (Kaneisha et al., 2004). We included in the annotation pipeline the analysis of non coding RNAs and we provided, for each tomato EST, a link to the identification numbers of TOM1, the reference cDNA microarray for Tomato (TED; http://ted.bti.cornell.edu/). The preliminary experimental annotation of the BAC sequences available from the SOL Genomics Network was based on the experimental data from tomato and potato ESTs and TCs (Traini and Chiusano, BITS 2006). The annotated BACs are available by the Generic Genome Browser (GBrowse) interface to allow selection for reference Gene Models to test predictive approaches. The two EST databases and the Generic Genome Browser are cross-referenced to provide an integrated platform.

Results
The platform was built to provide an italian resource for the genomics of Solanaceae family and for the International Tomato Genome Sequencing Project. The web based interface of the platform can be accessed browsing genome data in the form of BAC sequences or querying the EST resources from tomato and potato species. The platform is useful to support the experimental annotation of the genomic data, indeed, EST collections are a quick route for discovering new genes and for confirming coding regions in genomic sequences. Moreover, we believe that the integrated possibility to investigate on specific expression patterns as well as coding or non coding gene families from the annotated EST databases provide a relevant support for the investigation and the comprehension of genome organization and functionalities. We may well hope that our effort represent a framework for a useful organization of structure genomics data and for meaningful functional analyses based on comparative approaches with other model plant species to provide a reference within the Solanaceae community as well as for other similar efforts.

Availability: http://cab.unina.it/

Contact email: chiusano@unina.it

Supplementary informations

Acknowledgements This work is supported by the Agronanotech Project (Ministry of Agriculture, Italy).

SYMBiomatics: Synergies in Medical Informatics and Bioinformatics

Cameron G (1), Clark D (1), Beltrame F (2), Coatrieux JL (3), Del Hoyo Barbolla E (4), Martin-Sanchez F (5), Milanesi L (6), Tollis Ioannis G(7), Van der Lei J (8)

(1) EMBL-European Bioinformatics Institute, UK
(2) DIST University of Genova, Italy
(3) INSERM, France
(4) Ministry of Education and Science, Spain
(5) Institute of Health "Carlos III", Spain
(6) CNR-ITB - Institute of Biomedical Technologies, Italy
(7) Foundation for Research and Technology , Greece
(8) Erasmus Medical Center, Netherlands

Motivation

The European Commission has selected the EBI to coordinate a project that will stimulate and explore synergies between bioinformatics (the science of storing, retrieving and analysing large amounts of biological information) and medical informatics (the science of processing, sharing and using large amounts of medical information). The SYMBiomatics project will culminate in a White Paper that will inform the Commission's funding policy on the synergy between these two rapidly growing areas. The aim is to facilitate and accelerate biomedical research and innovation, with the ultimate goal of improving Europe's efficiency at developing better tools and systems for disease prevention, diagnosis and treatment. Building on decades of advances in deciphering the molecular components of living things, molecular and computational biologists are now synthesising the information that they've gathered, and are building a detailed understanding of cells, tissues, organs, organisms and populations. At the same time, clinical research has led to a better appreciation of the molecular basis of disease. Clinical scientists are amassing information that is helping them to decipher how variations in people's genetic make-up can affect their likelihood of developing certain diseases such as cardiovascular disease or diabetes, or of developing an adverse response to particular drugs, such as the anti-coagulants used to treat some types of heart disease. The development of technologies that will allow scientific and clinical information to be shared and integrated more readily will expedite the creation of novel diagnostic, preventive and therapeutic methods, allowing people to lead longer, healthier lives."

Results

Working together over the next fifteen months, an executive committee comprising nine organisations from six different European Member States (UK, France, Italy, Spain, Greece and the Netherlands) will document the state of the art in biomedical informatics. The group will identify areas of maximum opportunity, by systematically collecting insights from experts in the field and by analysing the scientific literature. Areas of opportunity will then be documented and prioritised. The group's findings will be presented at a meeting in early summer 2006, enabling further discussion by the wider community of bioinformaticians, medical informaticians, the growing number of clinical professionals whose work spans these domains and European policy makers. The project will culminate in a report that will summarise the project's findings and will provide input to future European scientific and funding policy.

Availability: http://www.symbiomatics.org/

Contact email: Dominic.Clark@genericsgroup.com

Microarray cross-platforms differential expression validation

Cordero F (1,2), Saviozzi S (2), Di Renzo MF (3), Olivero M (3), Cirenei N (4), Calogero RA (2)

(1) Dept. of Informatics, University of Torino
(2) Dept. of Clinical and Biological Sciences, University of Torino
(3) Laboratory of Cancer Genetics IRCC, Candiolo (TO)
(4) Applera Italia S.r.l (Monza)

Motivation

The microarray technology it is now a consolidated instrument for genome-wide analyses. Microarray exists in two different configurations: two channels microarray, i.e. two different species of RNA are investigated at the same time on the same array, and single channel microarrays, i.e. a single RNA specie is hybridized on each microarray. Many commercial platforms are now available as single channel arrays and cross-platform validation studies indicate that homogeneous results can be obtained independently by the platform used. Multiple microarray platforms analysis of the same expression data could be a powerful approach to large scale differential expression validation. Statistical tools used for differential expression validation, although optimized for microarray analysis, suffer of the limitations induced by the limited number of replicates. Especially type I error correction approaches, which is a critical step in the computational pipe-line for microarray analysis, is an important source of false negatives due to the limited efficacy of false-discovery estimation (Choe et al. Genome Biol. 2005, 6,R16). Furthermore, massive differential expression validation can not be afforded by quantitative RT real time PCR (qPCR), due to the high cost of experiments.

Methods

Three prototypic situations, derived by a time-course experiment designed to study the synergic effects of HGF and cis-platin on ovarian cancer, were used (Olivero et al. submitted): SK-OV-3 cells untreated cells (ctrl), SK-OV-3 cells treated with HGF for 48 hours (hgf) and SK-OV-3 cells treated for 24 hours with cis-platin (hgf/cddp) after 48 hours HGF pretreatment. Ctrl and hgf are available as biological triplicate as instead the hgf/cddp as duplicate. The three conditions were investigated using three different single channel platforms: Affymetrix hgu133av2, Illumina 24K arrays and Applera AB1700 Human arrays. Data where acquired with platform specific instruments and raw numerical data were normalized using the cyclic-loess approach.

Results

The availability of technical replicates on three different platforms, each characterized by some peculiarity (i.e. Affymetrix gene expression is derived by multiple different 25mer probes integration, Illumina gene expression is derived by the average of multiple replicates of the same 50mer probe, Applera gene expression is investigated using 60mers and a chemoluminescence detection system, as instead Affymetrix and Illumina use fluorescence detection) offers the opportunity to investigate the possibility to identify differential expression limiting the loss of information associated to type I error correction (Choe et al. 2005). On the basis of our observations we will discuss the possibility to design differential expression analysis pipe-lines based on data derived by at least two different single channel platforms.

Contact email: raffaele.calogero@unito.it

The bioPrompt-box: an ontology-based clustering tool for searching in biological databases

Corsi C, Ferragina P, Marangoni R

Department of Informatics, University of Pisa, Pisa.

Motivation

High-throughput molecular biology provides new data at an incredible rate; the increase in the size of biological databanks is enormous and very rapid. This scenario generates severe problems not only at indexing time, where suitable algorithmic technologies for data indexing and retrieval are required, but also at query time, since a user query may produce such a large set of results that their browsing becomes unaffordable. This problem is well known to theWebIR community and, in fact, third-generation search engines are currently providing new tools to better and better satisfy the "user needs" behind their queries: personalization (e.g. Eurekster), user behavior profiling (e.g. Google, Yahoo), query term suggestion (e.g. AskJeeves), are just a few of them. Web-results clustering is another approach pioneered by NorthernLight (1996) and then recently made famous by Vivisimo.com. The basic idea is that the results returned by a (meta-)search engine are clustered into folders which are labeled with intelligible sentences capturing the "theme" of the results contained into them. The folders are further organized in a hierarchy, whose internal nodes are labeled with sentences too. Users can therefore navigate through the labeled folders driven by the "need" behind their query. In this way (lazy) users are not limited to look at first ten results (more than 85% of web users do it), but they can immediately acquire several points of view on a larger pool of them. Consequently, they can either narrow their search by clicking on some folders, or they can acquire new knowledge by looking at the folder labels, and possibly refine their query. These nice features have driven some authors to say that "clustering technology is the future of search engines"

Methods

In this work we present a tool, TheBioPrompt-box, that follows the labeled hierarchical clustering approach in the context of biological databanks (specifically Uniprot, for now), thus exploring the applicability of this technology to the specialties of the biological data, with the ultimate goal of making the exploration task of biologists easier, more effective and more efficient. It goes without saying that, while Web pages are heterogeneous and uncontrolled so that their clustering must operate on-the-flight over the text excerpts returned by the queried search engines (cfr. Vivisimo.com); in the biological context, data are metatagged and come from multiple (humanly controlled) sources. As a result, the clustering and labeling task is, from one side, simplified by the availability of these metadata but, from the other side, it is made more challenging because it needs new methods for combining these sources into more effective labeled clusters. This is exactly the scenario we explore with TheBioPrompt-box which, at the best of our knowledge, is the first system in the literature adopting this kind of approach on biological data. In fact, ClusterMed is a similar project, but it offers a clustering only on the bibliographic database PubMed.

Results

At this stage, TheBioPrompt-box defines a meta-document as a sequence (protein or gene) plus the meta-data associated with that sequence in UNIPROT. So, a (meta)document is a protein sequence with the information about the organism, the comments of the researchers, the references to Gene Ontology, the references to articles or publications related to this sequence (possibly referring to Pubmed), or the taxonomy lineage of the organism. This means that the (meta) documents indexed by TheBioPrompt-box are composed by a list of fields containing either references to ontologies, or to external databanks, or they are plain text like researcher comments and (title of) articles. TheBioPrompt-box indexes this (meta) documents collection using Apache Lucene, a high-performance, full-featured and open-source text search engine library written entirely in Java, in

connection with Apache Commons Digester to manage XML format data. At query time, TheBioPrompt-box offers to the user some useful tools to customize the search and the clustering process. The more relevant query types currently supported are: Free Text over all fields of the (meta) document; Sequence-based which exploits Blast; and Keywords-based which searches over the keyword field of each indexed (meta) document. The search process is intended to select a set of (meta) documents over which the clustering algorithm is then executed. Currently, the user may cluster the (meta-)documents according to three different rules: the terms of Gene Ontology the (meta) documents cite, their lineage, or the organism information they share.

Availability: http://brie.di.unipi.it:8080/BioPrompt-box/

Contact email: marangon@di.unipi.it

Amino acid propensities for secondary structures

Costantini S (1,2), Colonna G (2), Facchiano AM (1,2)

(1) Laboratorio di Bioinformatica e Biologia Computazionale, Istituto di Scienze dell'Alimentazione, CNR, Avellino, Italy
(2) Centro di Ricerca Interdipartimentale di Scienze Computazionali e Biotecnologiche, Seconda Università di Napoli, Italy

Motivation
Propensity for secondary structures represents an intrinsic property of amino acid, and it is used for generating new algorithms and predictive methods. Our work has been aimed to investigate what is the best protein dataset to evaluate the amino acid propensities, either larger but not homogeneous or smaller but homogeneous sets consisting of only all-alpha, only all-beta or only alpha-beta proteins.

Methods
All analyses were performed using a set of experimentally determined, non-redundant protein structures in the PDB with mutual sequence homology <25%. The secondary structure for every PDB entry was assigned by the DSSP algorithm and was used to assign the secondary structural class, according to two definitions of structural classifications. The residue propensity values in different secondary structural types (Pij) were determined with an original software from the ratio of the residue's frequency of occurrence in helices, beta-strand and coil versus its frequency of occurrence in the PDB. An original prediction strategy was applied, based on the residue propensity for the secondary structures. For each protein amino acid, the average value of helix, beta strand and coil propensity has been determined by considering their surrounding amino acids in a window of length n. For each secondary structure, it has been evaluated the better window as well as the better coefficient to be applied to the average propensity of the segment. Finally, by comparing the average helix, beta strand and coil propensities for the segment under examination, the higher value was the criterion to assign the secondary structure of the amino acid in the middle of the segment. The quality of these predictions was examined by resubstitution and jackknife tests.

Results
We evaluated amino acid propensities for helix, beta-strand and coil in more than 2000 proteins from the PDBselect dataset. With these propensities, secondary structure predictions performed with a method very similar to the Chou and Fasman gave us results better than the original one, based on propensities derived from the few tens of X-ray protein structures available in the 1970s. Moreover, we subdivided the PDBselect dataset of proteins in three secondary structural classes, i.e. all-alpha, all-beta and alpha-beta proteins. For each class, the amino acid propensities for helix, beta-strand and coil, have been calculated and used to predict secondary structure elements for proteins belonging to the same class by using resubstitution and jackknife tests. The results were improved in comparison to the predictions for the whole PDB dataset (1). Therefore, amino acid propensities for secondary structures result more reliable depending on the degree of homogeneity of the protein data set used to evaluate them. Indeed, our results indicate also that all algorithms using propensities for secondary structure can be still improved to obtain better predictive results. We are developing a web-service to predict the secondary structure of proteins from their amino acid sequence using the amino acid propensities for secondary structures calculated for this study. The service will be available at the lab web server: http://bioinformatica.isa.cnr.it/

Availability: http://bioinformatica.isa.cnr.it/

Contact email: susan.costantini@unina2.it
angelo.facchiano@isa.cnr.it

References

1. Susan Costantini, Giovanni Colonna and Angelo M. Facchiano: "Amino acid propensities for secondary structures are influenced by the protein structural class.", Biochemical and Biophysical Research Communications (2006), 342, 441-451.

Systematic identification of stem-loop containing sequence families in bacterial genomes

Cozzuto L (1,2), Petrillo M (1), Silvestro G (3), Di Nocera PP (3), Paolella G (1,4)

(1) CEINGE Biotecnologie Avanzate, Napoli, Italy
(2) S.E.M.M. - European School of Molecular Medicine - Naples site, Italy
(3) DBPCM, Universita' degli Studi di Napoli Federico II, Napoli, Italy
(4) Dep. SAVA, Universita' del Molise, Campobasso, Italy

Motivation
Many bacterial genomes are known to contain families of repeated sequences of variable length and copy number. Many of them have been shown to contain a common stem-loop structure (SLS), involved in biological processes ranging from regulation of transcription to termination, to RNA stabilization. The availability of a now large number of sequenced bacterial genomes, represents an opportunity to identify novel families of such SLS containing repeated sequences. To carry out a systematic analysis of high stability stem-loop structures, predicted at both DNA and RNA level, an automatic pipeline was developed, based on Markov clustering algorithm (MCL). On each identified family, extensive annotations both at sequence and structure level were performed.

Methods
A pipeline has been previously described, able to identify, annotate and store into a relational database all potential SLSs within 40 completely sequenced bacterial genomes, representative of the bacterial world, from firmicutes to proteobacteria. From this population, SLSs predicted to fold with a free energy lower than -5 Kcal/mol were selected and filtered to eliminate those falling within tRNA/rRNA genes or known IS sequences. For each genome, SLSs were clustered according to a procedure based on BLAST and MCL programs (Altschul, S.F. and al. 1990 J. Mol. Biol. 215:403-410 and Enright A.J. et al Nucleic Acids Res. 2002 30[7]:1575-1584, respectively): an all-against-all SLSs BLAST comparison was performed for the creation of an e-value based distance matrix, which was in turn fed to MCL to produce a set of 'raw' clusters. Overlapping SLSs were fused into larger "regions". Members of the same cluster were aligned to produce a consensus sequence and the quality of the alignments was statistically evaluated. Genomic coordinates were used to classify clusters as either interspersed or tandemly repeated, and to join related raw clusters into the final refined set. BLAST analysis was used to identify clusters corresponding to repeated sequence families previously described in the literature. The families of identified sequences were analyzed to evaluate the reliability of the predicted secondary structure.

Results
Starting from a number of SLSs ranging from 6,534 (Buchnera) to 214,459 of (B. bronchiseptica) a first strand-dependent clustering step identified 717 "raw-clusters", each composed by a minimum of 7 regions. Clean-up and reclustering grouped raw-clusters in 466 second-order clusters. Clustering quality was assessed by alignment of the raw-clusters, which showed average identity above 60% for over 80% of the clusters. The established consensus was longer than 30 bp for over 98% of them. Further refinement led to a final set of about 85 clusters, including 28 known families and 29 families of not previously described repeated sequences. Analysis of predicted RNA secondary structure shows that many clusters are composed, only or predominantly, of SLSs with a low probability of random-folding ($p <= 0.005$), a clear enrichment over the original population, encouraging further studies aimed to better characterize their structure. A web interface is under development to allow public access to the SLS family database.

Contact email: cozzuto@ceinge.unina.it

TOMATEST DB: a database of expressed sequences to mine on Tomato functional genomics

D'Agostino N (1), Aversano M (1), Frusciante L (2), Chiusano ML (1)

(1) Department of Structural and Functional Biology, University 'Federico II', 80134 Naples, Italy
(2) Department of Soil, Plant and Environmental Sciences, University 'Federico II', 80055 Portici (NA), Italy

Motivation

TomatEST db is a secondary database with primary EST sequence information collected from Solanum lycopersicum (200,438), Solanum pennellii (8,346), Solanum habrochaites (8,000) and Solanum lycopersicum X Solanum pimpinellifolium (1,008) available at the dbEST in the release of November 2005. TomatEST can be considered a fundamental resource for the International Tomato Genome Sequencing Project (http://www.sgn.cornell.edu/help/about/tomato_sequencing.pl), since EST collections are a quick route for discovering new genes and for confirming coding regions in genomic sequences and a workbench for tomato functional genomics. One promising approach for the extraction of biologically meaningful information on genome functionalities is to find suitable way for the classification and the organization of the data. For the description of sequence function, we choose the Gene Ontologies (GO; The Gene Ontology Consortium 2000) and the Enzyme Commission (EC; Bairoch, 2000) numbers so to directly classify the annotated sequences according to their functionality and to link data to known pathways.

Methods

TomatEST db core structure is a MySQL relational database collecting all the results of the sequence analysis automatically produced by the execution of the software ParPEST (D'Agostino et al. 2005). TomatEST is integrated with two satellite databases: myGO and myKEGG. The first database is a mirror of the GO database. The second one is built from KEGG (Kaneisha et al., 2004) xml formatted files and related maps in GIFF format. The database provides a complete set of computationally defined transcript indices. The transcript indices represent unique sequences, singletons or tentative consensus (TC), derived from EST clustering analysis. The four different collections are clustered independently accounting for a total of 52,780 transcript indices, 18,813 TC and 33,967 singletons. We based the functional annotation, from both primary data (ESTs) and transcript indices, on BLAST similarity searches versus the UniProt database (Apweiler et al., 2004). In case of successful matches (e-value less equal than 0.01), the three best blast hits are stored into TomatEST db. When the UniProt identifier is in myGO db, the database is crosslinked to the gene ontologies and when the EC number occurs in the best blast hit description lines, the database is crosslinked with myKEGG. The TomatEST web site is a PHP-based interface that allows easy access to the data.

Results

TomatEST db can be queried through a pre-defined query system to support non expert users. All the results are displayed in detailed and friendly graphical views. The data can be queried via three different HTML forms. The first form produces an 'EST report page', the second form results in a 'Clusters report page', and the last form results in a 'Transcript Indices report page' which represents the core structure of the web interface. Here the data corresponding to user-selected criteria can be inspected considering two different classes of objects: the enzymes and the metabolic pathways. The results are reported as HTML based tree menus. This format allows immediate and informative data retrieval. The expressed sequences resulting from a query can be mapped 'on the fly' onto metabolic pathways that can be accessed as GIFF images.

Availability: http://cab.unina.it/

Contact email: chiusano@unina.it

Supplementary informations

Acknowledgements This work is supported by the Agronanotech Project (Ministry of Agriculture, Italy).

Identification of new putative mitotic genes through coexpression analysis

Damasco C (1) Grassi E (1) Ala U (1) Gatti M (2) Di Cunto F (1)

(1) Dipartimento di Genetica, Biologia e Biochimica dell'Università di Torino, Via Santena 5 bis - I-10126 Torino, Italy
(2) Dipartimento di Genetica e Biologia Molecolare, Istituto Pasteur-Fondazione Cenci Bolognetti, Istituto di Biologia e Patologia Molecolari del CNR, Università di Roma "La Sapienza", P.le A. Moro 5, 00185 Roma, Italy

Motivation
Microarrays are one of the most powerful techniques for the analysis and the measurement of gene expression. In the last few years, a huge amount of expression data, obtained with this technology from several model organisms has become available. Therefore, the development of processing tools aimed to extract useful functional information from published primary data has become a very important computational biology subject. We have previously described a data-mining approach called CLOE (Pellegrino et al., BMC Bioinformatics 2004) based on meta-analysis of microarray datasets from pairs of species, which evaluates genes coexpression and its phylogenetic conservation among species. With this approach, it's possible to make high confidence predictions about proteins' function and interaction. Since the approach can be applied to any couple of species, we have investigated the possibility of massively using it to identify new putative mitotic genes. In particular, we concentrated on Drosophila Melanogaster, which allows high-throughput experimental validation by RNAi. Furthermore we are exploring the possibility of using coexpression analysis for identifying the best positional candidate gene for mitotic mutations mapped to wide genomic loci by classical genetic approaches.

Methods
DNA microarray have been obtained from published studies performed with both Affymetrix (De Gregorio et al., PNAS 2001) and cDNA platforms. In particular, the data on which the best positional candidate predictor is based and the mammalian data used for combined analysis have been collected from the Stanford Microarray Database (http://smd.stanford.edu/cgi-bin/search/QuerySetup.pl). All the probes were completely re-mapped to the most significant biological databases, and in particular, Unigene, Entrez Gene and Ensembl. The orthology relationships were assigned on the base of HomoloGene tables. The computational analysis starts with the selection of database's probes referred to genes that compose a locus. For every probe we generate a list of all the other dataset's probes, ordered by a coexpression index. To this regard, we use the Pearson's correlation coefficient. For the functional statistical analysis, we introduce annotations of the Gene Ontology (GO) project. The basic mechanism of the predictor is to assign to every gene a functional score, based on the ranks of the genes annotated to the relevant GO function in the corresponding coexpression list. To gain a further level of confidence, the analysis is then performed on the orthologous genes of other species, such as human and mouse.

Results
The systematic use of our approach has led to the identification of a huge number of candidate Drosophila mitotic genes. For many of them, the prediction has been successfully validated by RNAi, leading to the discovery of new mitotic functions. Besides to this results, we will present the preliminary description and validation of our best candidate prediction tool.

Contact email: ferdinando.dicunto@unito.it

Significance analysis of microarray transcript levels in time series experiments

Di Camillo B, Toffolo G, Cobelli C

Information Engineering Department, University of Padova, 35131 Padova, Italy

Motivation

Microarray time series studies are essential to understand the dynamics of biological molecular events. In order to limit the analysis to those genes that change expression over time, a first necessary step is to select differentially expressed transcripts. This is often accomplished using an empirical or statistically based fold change threshold and comparing samples time by time. This approach is far-from-ideal since it does not account for the dynamic nature of the data and is particularly sensitive to random fluctuations due to the noise. To overcome these limitations a variety of methods were proposed; among them ANOVA based procedures, run test, autocorrelation and approaches based on regression modeling. However, these methods are seldom applicable in practice since they require either a large number of replicates or a relatively high number of time samples, while microarray time series experiments usually consist of a limited number of samples and replicates are only available for a limited number of them. Here we present a novel algorithm to select differentially expressed genes, which accounts for the entire dynamic profile and explicitly handles the experimental error. The method requires a relatively small number of replicates, which makes it applicable for time series microarray analysis.

Methods

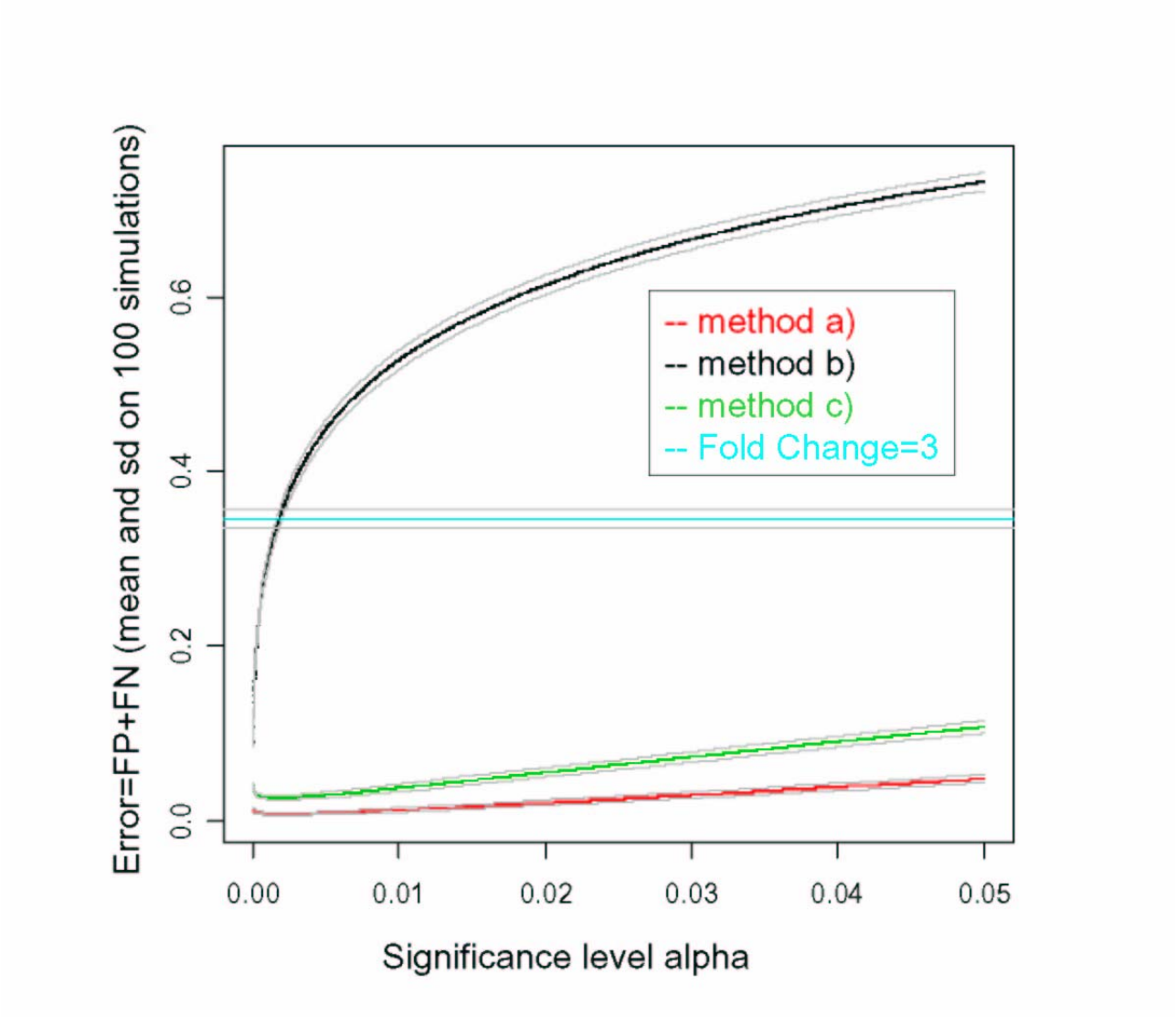
Three approaches to select differentially expressed genes are tested and their performances are compared using synthetic data: a novel method (a) and two methods proposed in the literature (b and c). Approach a) Let's call xT(tk) and xC(tk) the log-expression measurements in treated (T) and control (C) experiments, available for a generic gene x at time sample tk (k=1, ..., M). The rationale adopted to decide whether a gene x is differentially expressed in condition T and C is to calculate the area Ax of the region bounded by the two profiles and to consider it significantly different from 0 if it exceeds a threshold level D. For a given significance level alpha, the threshold D is derived from the distribution of the variable Ax when xT(tk) and xC(tk) are experimental replicates, i.e. the difference in the measurements represents only experimental variability. To this purpose Ax is calculated using a Monte Carlo approach for 10000 random time series, obtained by sampling the distribution of the following variable: dk=xT(tk)-xC(tk) derived when xT(tk) and xC(tk) are experimental replicates. Approach b) A threshold is derived based on the error distribution to be applied to dk for each time sample; genes are selected as differentially expressed if dk exceeds the threshold in at least one comparison. Approach c) Time series are fitted using a quadratic model of time. Genes are selected as differentially expressed if the differences between parameters estimated in T and C are significantly different from 0. Synthetic data 100 synthetic data-sets are generated, each consisting of 2000 genes and 10 time samples. In each data set 100 time series are generated using a Markov model in which dk+1 depends on dk according to a given probability model based on real data observations (multiple sets of differentially expressed genes), while 1900 time series are generated as random noise, modeled as a mixture of Gaussian as observed from real data.

Results

Method a outperforms methods b and c both in sensitivity and specificity, over a wide range of significance levels alpha. In Figure 1 the error (sum of false positives and false negative classifications) is reported as average (+/- standard deviation) on 100 simulations, for alpha ranging from 0 to 0.05. Results obtained using a constant fold change method (not depending on alpha) for each time sample, and selecting a profile as differentially expressed if it is differentially expressed in at least one time sample are also reported.

Availability: http://www.dei.unipd.it/~dicamill/Abstract

Contact email: dicamill@dei.unipd.it



3D-protein C mutation database: integration of structural, functional and clinical data of natural protein C mutants

D'Ursi P (1), Marino F (2), Caprera A (2), Milanesi L (2), Faioni E (3), Rovida E (2)

(1) Department of Science and Biomedical Technologies, University of Milano, Italy
(2) Institute of Biomedical Technologies, National Research Council, Segrate (Mi), Italy
(3) Hematology and Thrombosis Unit, Ospedale San Paolo, University of Milano, Italy

Motivation

Protein C (PC) is a vitamin K-dependent anticoagulant plasma serine protease that exerts its action through the inactivation of factors Va and VIIIa. In addition it plays an important role in inflammation and cell proliferation. Several mutations of the PC gene have been found in patients with protein C deficiency, a condition that is associated to the risk of developing venous thrombosis. PC deficiency determine a reduced plasma concentration and/or catalytic activity of the protein. In previous work, we have identified 33 mutations (18 novel) in the promoter and coding regions of the PC gene by PCR and sequencing in 46 patients reporting venous thromboembolic events. We have constructed the molecular models of missense mutations localized in the structurally resolved regions of PC, starting from PDB x-ray coordinates (pdb ID: 1aut) and have performed detailed computational analysis of those presenting aminoacidic substitution in critical positions for structure and function. The availability of structural models can be useful in the research and clinical fields to elucidate how a mutation may interfere with enzymatic activity, ligand binding and cofactors interaction and relate the effect to patient phenotype. Thus, we started to collect our data in a database that was conceived to be freely available for consultation and data retrieval. The database was then enriched with already described variants extracted from other sources (literature, SwissProt, HGMD). Our effort was to create an updated interactive tool to integrate clinical and phenotypical descriptions with functional and structural data obtained by computational approaches to help elucidate the chain of events leading from a molecular defect to pathology.

Methods

MySQL database management system and PHP web programming language were used to construct database and graphical interface. The database consists of 4 tables where all the informations about protein variants are stored. It contains 197 entries that include 184 missense and 13 stop mutations. Multiple alignments were obtained with CLUSTALW using a set of orthologous and paralogous sequences to show the residue conservation between species and within serine protease family. A in-house developed Perl script was used for process the CLUSTALW output, automatically producing for each variant an HTML-format multi-alignment file with the mutation highlighted in red. Molecular models were constructed manually in InsightII (Accelrys) or automatically with a modified script in Python of Modeller (Andrej Sali), both methods consisting in residue replacement and the conformation of the mutant side chain is optimized by coniugate gradient and refined using molecular dynamics for Python script. Structural models can be visualized as real-time 3D images by using RasMol and VRML (Virtual Reality Modeling Language). The functional and mutation sites are automatically mapped on the structure and highlighted with different colors. The output format for VRML was generated by MolScript v2.1.2 from the stored PDB coordinates.

Results

We have realized a specialized relational database and a search tool for natural mutants of Protein C. A query page allows the user to retrieve entries by position in sequence of a mutated residue, by aminoacid substitution, by keyword and by domain localization. The query results are listed in a table where each entry is linked to a details page. This page resumes data about gene, secondary structure and domain localization of the mutated residue. A structural data section reports multiple alignment files which highlights the substituted position and help to evaluate the degree of conservation, the coordinates of modelled variants, and a gallery of 3D images that illustrates the

structural implications of the aminoacidic substitution as long as the results of computational analysis, when available, like electrostatic potential representations and molecular dynamics trajectories. Clinical, phenotypic and functional annotations, manually extracted from the literature, are also reported. Direct links to relevant literature references and, when present, to the corresponding SwissProt Variant Page are included. The site provides tools to analyze and mutate new variants, not included in database. Given a position in sequence, it is possible to generate a multiple alignment highligthing the residue of interest in homologs and to mutate the protein using the program Modeller (A. Sali & T.L. Blundell.J. Mol. Biol. 234, 779-815, 1993) through an implemented Phyton script. 3D-protein C mutation database can be accessed at the site http://www.itb.cnr.it/procmd/ Acknowledgments This work was supported by European Project BioinfoGRID (Bioinformatics Application for Life Science).

Availability: http://www.itb.cnr.it/procmd/

Contact email: ermanna.rovida@itb.cnr.it

Immunogrid - The European Virtual Human Immune System Project

Emerson A, Rossi E

High Performance Systems Division, CINECA, via Magnanelli 6/3, 40033 Casalecchio di Reno (BO), Italy

Motivation

The immune system is a complex and adaptive learning system which has evolved to defend the individual against foreign invaders. It has multiple levels (molecular, cellular, organ and tissue, organism, and organism-to-organism) and is also combinatorial in nature with a large number of products; there are typically > 10^15 antibodies and 10^12 immune system cell clones in a single individual. The function of the immune system depends on both the genetic composition and the previous exposure, i.e. the experience of the organism. Immune intervention, such as vaccination, is the most effective method for the control of disease and the greatest achievements include eradication of smallpox, near-elimination of polio, and savings of some 170 million person-years. Vaccination has been used in the control of over two dozen diseases by the 50 or so successful vaccines which have been developed to date. These vaccines largely protect against infectious diseases, although recent vaccine developments offer great hope for treatment for a broader range of diseases. Large-scale studies of the immune system, also known as immunomics, is the key factor driving the current wave in vaccine development. These include genomics and proteomics, analysis of the diversity of pathogens or complexity of the human immune system, high-throughput screening or immunoinformatic tools for the management and analysis of vast quantities of data. Computational models are becoming increasingly important in immunomics: Experimental approaches are expensive and it is impossible to perform systematic experimental studies of immune processes in humans. Because of ethical issues, there are stringent limitations as to what experiments can be performed in humans. The usefulness of computational approaches to the study of immune system has been demonstrated, but computational models that encode the natural-size immune system have not been developed because of the past limitations of computational infrastructures.

Methods

In order to overcome the limitations of current immune system models the ImmunoGrid project was created. This initiative is a three year project which has been funded by the European Union and involves a consortium of leading European institutions from Italy, France, UK, and Denmark and also from Australia. The outcome will be a Virtual Human Immune System simulator that can be used as a computational tool for preclinical/clinical applications of vaccine development and immunotherapy. The project will address all aspects of immune system models including integration of standardization concepts and information on molecular, cellular and organ levels for the description of immune system processes and function, while the simulator itself will be validated by pre-clinical mice models. The key feature though will be the use of Grid technology to provide the necessary computational resources (CPU time and data storage) in order to cope with the natural complexity of the human immune system.

Results

The project officially started on the February 1st 2006 and has begun by addressing improvements in the current simulation models. A prototype of the Virtual Human Immune Simulator will be available within 18 months of the start of the project. The set of tools developed will be validated with experimental data and then provided to support clinical applications for the development of immunotherapies in cancer and chronic infections and disseminated to users such as vaccine and immunotherapy researchers and developers.

Availability: http://bioinfo.cineca.it/immunogrid

Contact email: a.emerson@cineca.it

Dynamic Regulation of Gene Expression

Farina L (1,2), Marcatili P (3,4), Uva P (3,4), Busiello V (1,4), Morelli G (1,3), Ruberti I (1,4)

(1) Molecular Systems Biology Laboratory, Rome, Italy
(2) Department of Computer and Systems Science "Antonio Ruberti", University of Rome "La Sapienza", Rome, Italy
(3) National Research Institute of Food and Nutrition, Rome, Italy
(4) Institute of Molecular Biology and Pathology, National Research Council, Rome, Italy

Motivation
One key point in the analysis of gene expression dynamics is that mRNA abundance is determined by two regulated processes: transcription and degradation, both specifically affecting transcript levels. It is becoming progressively more evident that mRNA degradation and its regulation is an important factor in determining the expression pattern of many genes. When trying to infer the global phenotypes of cells from large-scale mRNA expression profiling data, it would be important to consider both transcriptional and post-transcriptional level of gene regulation.

Methods
Here we model gene expression in order to capture the specific part of mRNA expression dynamically changing in response to regulatory signals affecting transcription, mRNA stability or both. To this end, we introduced the net mRNA expression, i.e. the mRNA expression normalized to basal transcription and degradation rates, computed by the NEMES (NEt MRNA ExpreSsion) algorithm.

Results
Evaluating net-mRNA expression using genome wide cell cycle time series data and decay the method strikingly identifies regulatory switching points, thus revealing the underlying dynamic link between gene expression and regulatory programs. As global decay data can be easily obtained by means of microarray experiments, we anticipate that net-mRNA expression will prove to be a powerful tool to fully understand the dynamics of gene regulation in single and multicellular organisms.

Availability: http://www.dis.uniroma1.it/~farina/NEMES/

Contact email: lorenzo.farina@uniroma1.it

A computational approach for detecting peptidases and their specific inhibitors at the genome level

Fariselli P (2), Bartoli L (2), Calabrese R (2), Mita DG (1), Casadio R (2)

(1) Department of Experimental medicine, University of Naples Federico II, Naples, Italy
(2) Laboratory of Biocomputing, CIRB/Department of Biology, University of Bologna, Bologna, Italy.

Motivation
Peptidases are proteolytic enzymes essential for the life of all organisms and responsible for fundamental cellular activities, such as protein turn-over and defense against pathogenic organisms. It is known that in Eukaryots about 2-5% of the genes encode for peptidases and peptidase homologs irrespectively of the organism source. The basic protease function is however "protein digestion". This activity can be potentially dangerous in living organisms, if not strictly controlled and for this reason, in vivo several specific inhibitors are present. Four major classes of peptidases are identified by the catalytic group involved in the hydrolysis of the peptide bond: 1) Serine Peptidase; 2) Aspartic Peptidases; 3) Cysteine Peptidase; 4) Metallopeptidase. Here we give a solution to the following basic problems: 1) recognize in human and mouse genomes both protease and protease inhibitor sequences; 2) predict which inhibitor is specific for a given peptidase. More specifically we address the following questions: 1) Given a pair of sequences, are they a pair of protease and inhibitor that can interact? 2) Given a protease (or inhibitor), how can we compute the list of the proteins in a defined database that can inhibit (or be inhibited by) the query protein? 3) Given a proteome, how can we compute the lists of peptidases and their relative inhibitors for each of the protease class described above?

Methods
In the last years, an invaluable source of information about proteases and their inhibitors has been made available through the MEROPS database, so that it is possible to look for peptidase or peptidase-inhibitor sequences (or structures). We tested PROSITE and PFAM and we integrate them in a unique framework taking advantage of their different behavior in detecting proteases and their inhibitors. We analyze the four major protease classes (Serine, Cysteine, Aspartic and Metal) and we test PROSITE and PFAM accuracy in the detection of proteases and inhibitors with respect to a non redundant set of globular proteins. In order to predict whether pairs of peptidases and inhibitors belong to the same class, we developed a system that performs two consecutive tasks: 1) extracts protease and inhibitor sequences from a given data set and labels them as belonging to one of the 4 classes; 2) tests whether the inhibitor can interact with the protease (yes, when the two sequences belong to the same class, eg serine peptidase and serine peptidase inhibitor). In order to address this problem, we designe a decision-tree method that processes the information obtained from PROSITE and PFAM and detects whether a query sequence can be classified as peptidase or inhibitor. The decision tree first uses PROSITE scan the data base in order to extract a protease and an inhibitor. When it fails the decision tree switches to PFAM.

Results
We first test PROSITE in the task of detection proteases and protease-inhibitors against a non redundant set of globular proteins that does not contain them. The accuracies are 70 and 90% for the proteases and inhibitors, respectively. Interestingly the number of false positive is nearly 0. This indicates that PROSITE has a very high specificity. When PFAM is scored using the same data sets the accuracies become 95 and 97% for the proteases and for the inhibitors, respectively. When the decision tree is adopted accuracies further improve of one percentage point (96% and 98%), suggesting that PROSITE specificity increases PFAM accuracy. However, in order to be able to measure the real system accuracy in the task of selecting pairs of proteases and interacting inhibitors, we compute the score of detecting pairs of protease-inhibitor among all possible pairs, namely peptidase/inhibitor, peptidase/other, inhibitor/other, peptidase/peptidase, inhibitor/inhibitor, other/other, excluding the self-combinations (a sequence against itself). Proceeding in this way we

ended up with a number of 18.559.278 pairs. We divided peptidase sequences in the four classes according to their biological activity as detailed above. We labeled the inhibitors accordingly, with the exception of one more class (U) containing inhibitors that are reported to be able to inhibit to some extent all types of peptidases (Universal inhibitor). Out of all the possible 18.559.278 pairs only the ones that pertains to proteases and inhibitors of the same class are counted as members of the positive class (true positive in the confusion matrix), which amounts to less than 7% of the all possible pairs. Scored on this stringent data set, the decision tree method achieves a joint global accuracy of 94% with a coverage for the positive cases (protease-inhibitor) of 99%.

Availability: http://www.biocomp.unibo.it/

Contact email: remo@biocomp.unibo.it

Alpha and Beta Estrogen Receptors: Molecular Modelling and Conformational Analysis through Molecular Dynamics

Ferrario MG (1,2), Chiappori F (1,2), Gaiji N (1,2), Fantucci P (1,2)

(1) Department of Biotechnologies and Biosciences, University of Milano-Bicocca, Milano.
(2) DELOS s.r.l, Bresso (MI)

Motivation

Estrogens, a most important group of steroidal hormones, regulate sexual differentiation and functions, preside bone construction, remodelling and homeostasis, lipoproteins synthesis, learning and memory functions. They protect the central nervous and cardiovascular system from the risk of degenerative diseases. In women, after menopause, these functions are seriously impaired; as a consequence, it is of deep interest to find a compound which could substitute the endogenous estrogens without eliciting the dangerous side effects that seem to be associated with the traditional Hormone Replacement Therapy (HRT), e.g. mammary and uterine cancer. In this view, it seems important to design compounds, of steroidal and non-steroidal origin, that can discriminate between the two human Estrogen Receptor (ER) isoforms: ERα and ERβ. Actually, the ERβ isoform seems to be principally correlated with the non-sexual functions of the estrogenic compounds, while the ERα form is suspected to be more correlated to the cancerogenic side effects of HRT. These are the reasons for the present investigation of the dynamical behaviour of the ERα- and ERβ-LBD models in explicit water, in different conformations (agonist and antagonist), in the presence or absence of their endogenous agonist ligand 17-β-Estradiol (E2). This was achieved by Molecular Dynamics (MD) protocols and Essential Dynamics (ED) analysis.

Methods

The starting structures of ERα and ERβ Ligand Binding Domains (LBD) were 1G50[1] (α-LBD bound to E2, agonist, at 2.90Å) and 1ERR[2] (β-LBD bound to RAL, antagonist, at 2.60Å). The lacking β structures, in agonist (BsuA) and antagonist (BsuAa) conformation, have been generated by Homology Modelling using the Swiss-Model server provided with the template of the corresponding α structure, the ERβ-LBD sequence was derived from the 1QKM[3] PDB structure. E2 was then rigidly docked to 1G50 and BsuA models, using the AutoDock package. The models were then optimized through a Molecular Mechanics (MM) protocol of energy minimization, through which the ER-LBD structure was released gradually. 1.MM on side chains with backbone frozen, in vacuo. 2.MM with the whole protein frozen, in water. 3.MM on the side chains with the backbone frozen, in water. 4.MM on the whole system, protein and water. The constraints applied to the ligand were the same as those applied to the protein side chains. This optimization was performed by the GROMACS package. The resulting models were submitted to the PDB-Procheck package, ADIT, to check their structural consistency. Six models were optimized: 1G50, BsuA, 1ERR, BsuAa, all apo-receptors, and 1G50-E2 and BsuA-E2, complexes. On these structures NVT-MD simulations have been performed using the GROMACS package, after a 60ps thermalization during which the temperature was gradually increased to 300K. The MD runtime has been of 12ns for 1G50, BsuA, 1ERR, BsuAa and 3ns for 1G50-E2, BsuA-E2. A specific topology file for E2 has been built for the GROMACS force field, as non standard data exist for such a ligand. These simulations were analysed by the ED method to point out the principal components (PCs) of the molecular motion.

Results

All the simulation trajectories resulted to be reliable and significant as indicated by the cosine content test (cc), the cc is a measure of the similarity of the trajectory to a random diffusion. BsuA model was found to be the most flexible structure, interacting colser with the ligand, whose presence reduced drastically the structure mobility. 1G50 was found to be more loosely bound to the E2 ligand, so that the ligand presence influenced less severely the structure. As expected, the maximum flexibility of the structure is mapped in the peripheral loops of the protein, but it is

differently distributed in each system. The accessibility of the binding site got reduced along the simulation in the absence of the ligand, while it kept constant in its presence. The presence of the ligand, which is essentially non-polar, causes an increase in the hydrophobic surface area accessible to the solvent, in relation to the hydrophilic one. The hydrogen-bonding network of the ligand-protein interaction, was differently characterized in 1G50 and BsuA. In the first case, the number of H-bonds varies from zero to four, while in the second case one or two bonds are present during the whole simulation. This correlates well with the observed minimum distance between the ligand and the protein. The behaviour of the models is quite well parted between the agonist and antagonist conformations, so that, from the point of view of motion, 1G50 and BsuA could be clustered together, as well as 1ERR with BsuAa. This difference in behaviour is particularly enhanced in the values of the parameters characterising the H12. Actually, these topological parameters are conserved along the simulation in the case of antagonist conformations, while rigid movements are wider for agonist conformations.

Contact email: noura.gaiji@unimib.it

References

1. Eiler et Al., Protein Expression Purif. 2001.
2. Brzozowski et Al., Nature 1997.
3. Pike et Al., EMBO J. 1999.

A novel structure-based encoding for machine-learning applied to the prediction of SH3 domain specificity

Ferraro E, Via A, Ausiello G, Helmer-Citterich M

Department of Biology, University of Tor Vergata, Roma

Motivation
Protein recognition modules (PRMs) play a key role in the frame of protein-protein interactions. PRMs are protein domains that focus their binding targets on short protein sequences of about ten residues. SH3 domains are well-studied PRMs that bind proline-rich short sequences characterized by the PxxP consensus. The binding information is typically encoded in the conformation of the domain surface and in the short sequence of the peptide. We extracted this information as significant pair of residues involved in the interaction and defined a numerical encoding scheme in order to build a predictive model for the SH3 domain specificity.

Methods
In the first step of the methodology, pairs of contact residues between an SH3 domain and a ligand-peptide are identified. The contact information is extracted from SH3-peptide complexes of known structure and can also be derived for complexes whose structure is unknown, but can be built with homology modeling techniques From pep-spot and phage-display experiments (Landgraf C. et al., (2004) PLOS Biol. 2, 94-103, Brannetti B & Helmer-Citterich M., Nucleic Acids Res. 2003 Jul 1;31(13):3709-11), we obtained a dataset of interacting and non-interacting domain-peptide pairs. Binding information is then numerically encoded and used to train a neural network. The encoding procedure is based on the frequency of contact residues characterizing binders and non-binders. The neural model prediction is given in terms of the binding propensity of a domain-peptide pair. We focused our analysis on a group of sixteen yeast SH3 domains and a library of 780 peptides from the yeast proteome, resulting in 8797 domain-peptide pairs. Of such pairs, 649 are interacting and 8148 are non-interacting. The method is applicable to other organisms and even to other families of PRMs, whenever at least one complex of known structure and some experimental data on domain-peptide interactions are available. To verify the generalization capability of the model we scanned the peptide interaction library of the human SH3 domain Abl.

Results
The model was tested with five-fold cross validation on the entire yeast dataset. The results are very encouraging: the global accuracy of the model, measured by the area under the ROC curve (AUC), exceeds 80%.

Contact email: enrico@cbm.bio.uniroma2.it

.

Bayesian approaches for reverse engineering of cellular networks: a performance evaluation on simulated data

Ferrazzi F (1), Sebastiani P (3), Kohane IS (2), Ramoni MF (2), Bellazzi R (1)

(1) Dipartimento di Informatica e Sistemistica, Università degli Studi di Pavia, Italy
(2) Children's Hospital Informatics Program and Harvard Partners Center for Genetics and Genomics, Harvard Medical School, Boston, USA
(3) Department of Biostatistics, Boston University School of Public Health, Boston, USA

Motivation
Time series measurements of gene expression or protein concentrations allow us to model the cell as a dynamic system, whose instantaneous state can be characterized by a set of state variables. Dynamic Bayesian networks (DBNs) are a special class of BNs particularly suited to study dynamic data. In particular, Linear Gaussian Networks allow researchers to avoid information loss associated with discretization and render the learning process computationally tractable even for hundreds of variables. However, it is often argued that linear models cannot capture the complex nonlinear dynamics of cellular systems. For this reason, we here propose a model that uses a linear regression of nonlinear transformations of the parent values. We evaluate both approaches using simulated data produced by a mathematical model of cell cycle control.

Methods
Assuming to have a database of measurements for n genes/proteins in p consecutive time points, it is possible to derive the DBN which encodes the dependencies over time of the random variables representing gene expression or protein concentration values. Supposing that the process under study is first order Markovian and stationary and that no instantaneous relationship between the values of two variables is possible, we need to learn only the transition network between the variables at time t and at time t + 1. To this aim, a probability model and a search strategy must be chosen. Linear Gaussian Networks treat variables as continuous and exploit a linear regression model to describe the conditional mean of a variable with respect to its parents. We here propose also a model in which this conditional mean is instead described as a linear combination of nonlinear functions of the parents. As it is reasonable to assume that expression/protein levels cannot indefinitely grow in proportion to their parent values, we decided to use the hyperbolic tangent function, in order to model a saturated effect of the parent on its child. In accordance with the Bayesian literature, we look for the network with maximum posterior probability given the data: to this aim, we exploit a finite horizon local search and we explore the dependency of each variable on all the variables at the previous time point.

Results
In order to have some insights regarding the suitability of Gaussian networks to describe relationships among genes or proteins, we decided to carry out an experiment with simulated data coming from a model of the budding yeast cell cycle (Chen et al., Mol. Biol. Cell, 2004). The whole model contains 36 differential equations: almost all the 36 variables represent protein concentrations, while the others represent the mass and the timing of cell cycle events. We used the profiles simulated in the case of wild-type cells and sampled values every 5 mins, from time 0 to 100 mins (about one cell cycle length). We analyzed the obtained dataset with the proposed dynamic Bayesian network approach, using both the linear regression model and its modification with nonlinear functions. If we consider as "true parents" of a variable A the other variables that appear in the differential equation describing A's dynamics, we can compare these with the parents found by the DBN algorithm. It is therefore possible to calculate the recall and precision: the recall corresponds to the fraction of "true parents" correctly inferred by the DBN algorithm, while the precision is the fraction of inferred parents that are also "true parents". However, considering only the parent variables in the regulatory network can constitute a very restrictive criterion: the value of a variable A at time t is in fact at the same time influenced by its parent values at time t - 1 and an

important determinant for its children values at time t + 1. We therefore decided to repeat the accuracy calculations comparing the "true parents" of each node with the variables in its Markov blanket, given by the union of its parents, its children and the parents of its children. The recall obtained with the nonlinear function model is always slightly higher and the precision slightly lower than the ones obtained with the linear model. On average both models provide results characterized by a 30% recall and an analogous precision. The inferred networks are generally very parsimonious (each variable has few parents) and statistical analysis showed that the goodness of fit is satisfactory. These preliminary results confirm the suitability of Gaussian networks for a first level, genome-wide analysis of high throughput dynamic data: the models here proposed can indeed infer a synthetic and quite accurate description of the system under study, useful to guide researchers to further, deeper studies.

Contact email: fulvia.ferrazzi@unipv.it

Mining the human interactome through gene expression time series analysis

Ferrè F (1), Clote P (1), Ausiello G (2), Via A (2), Cesareni G (3), Helmer-Citterich M (2)

(1) Biology Dept., Boston College, MA
(2) Center for Molecular Bioinformatics, Dept. of Biology, University of Rome Tor Vergata, Rome
(3) Dept. of Biology, University of Rome Tor Vergata, Rome

Motivation

Gene expression as measured by DNA microarray experiments offers at a glance an overall view of cellular processes, thus revealing complex relationships among genes. Sampling gene expression levels at different points in time produces a dynamic landscape of expression changes. This offers unique perspectives for the clustering of genes with similar temporal expression and for the reconstruction of functional modules. Time series microarray experiments are ideal for the study of dynamic processes, such as cell cycle, development and changes in gene expression in response to different levels of a new condition, like a drug, and can be used to highlight differences and similarities between different tissues, or between normal and pathological conditions. On the other hand, protein interaction networks can provide information about the skeleton of physical interactions underlying the relationships among the involved genes. Our goal is the study of human interaction networks and their temporal evolution in different tissues and conditions by means of the clustering of interacting proteins with similar temporal expression. The integration of gene expression time series data with interaction networks is a powerful tool for the discovery of gene modules and the analysis of how these modules are perturbed in pathological conditions, thus creating the basis for a molecular pathology of diseases.

Methods

While algorithms for the analysis of static microarrays can in principle be applied to dynamic microarray data, the study of gene expression time series raises specific problems and requires specifically developed computational tools in order to fully exploit the data information content. Dynamic time warping (DTW) is an algorithmic technique for the computation of the smallest distance and optimal alignment between two numerical sequences. DTW is evidently a very flexible tool for time series data comparison [1] since it identifies similarity between sequences where there is a shift in the time axis, and it accommodates sequences of different length. The quality of an alignment is estimated by the time warping distance (TWD), which is an alternative and possibly better measure than the commonly used distance measures (Pearson correlation coefficient, Euclidean distance, L1 distance). We used TWD as distance measure for our implementation of the Cluster Affinity Search Technique (CAST) algorithm [2], which is of particular interest for the analysis of noisy data since it explicitly incorporates an error model, thus it is particularly appropriate for biological applications. Unlike most clustering algorithms, CAST makes no assumptions about the number of clusters, their size or structure, which are discovered from the data. Arguably, the most relevant property of DTW is that it can identify and align sequences, which have approximately the same overall component shapes, but these shapes do not line up along the x-axis. DTW warps the time axis of one or both sequences to achieve a better alignment. This time shift may be indication of a functional relationship where the expression of one gene activates one or more other genes. Clustered genes can identify a gene module where similar expression through time is an indication of functional interaction. Genomic data helps the pathway reconstruction, for example through the identification of similar transcription factors binding sites. Sub-cellular localization (known or predicted) is used to further filter the clusters.

Results

Human interactome modules, obtained by the clustering of public human time series data [3-5], are benchmarked using curated protein-protein interactions from the MINT database [6] and gene ontology (GO) annotations, and a confidence index is derived accordingly. Cluster of genes

similarly expressed through time are identified, highlighting gene modules which are associated with specific functions (for example cell cycle progression regulation), or which are affected in pathological conditions.

Contact email: citterich@uniroma2.it

Publime: a new tool for meta-analysis of cancer-related microarray experiments

Finocchiaro G (1,2), Mancuso F (1,2), Muller H (1,2)

(1) European Institute of Oncology, Milan, Italy
(2) IFOM, Firc Institute of Molecular Oncology, Milan, Italy

Motivation

Application of gene expression microarray technology is rapidly producing large amounts of data that represent a considerable resource to produce a general view of the transcriptional activity of several organisms, both in different phases of development and in different conditions. The analysis of a specific microarray experiment profits enormously from cross-comparing to other experiments; statistical techniques of meta-analysis have been produced to integrate and to compare raw data generated from several microarray experiments. Nevertheless the way of dataset selection for meta-analysis represents a serious limitation for the identification of new and unexpected connections between different datasets. To facilitate the solution of these problems, we developed Publime (acronym for PUBlished LIsts of Microarray Experiments), a dedicated database, where researchers can find and consult published gene lists derived from gene expression microarray analysis.

Methods

Lists of genes are manually extracted from publications concerning microarray studies by experienced curators. Each gene is annotated following a procedure similar to those we adopted to generate IFOM Affymetrix annotation tables [1]. The annotation process is necessary both to synchronize information with latest database releases and to standardize heterogeneous identifiers found in publications. Whenever available, we associated each identifier with UniGene ID, NCBI Entrez Gene ID, Official Symbol, UniGene Title, Chromosome, and GeneOntologies. Medline annotation of each publication inserted is downloaded in XML format through NCBI Entrez Utilities http://eutils.ncbi.nlm.nih.gov/entrez/query/static/eutils_help.html via the perl LWP package. In particular, we extract the following information for each publication: Authors, Title, Abstract, Journal, Medical Subject Headings and Chemical that represent the controlled vocabulary of biomedical terms and chemical terms, respectively, used for indexing documents in MEDLINE. Data regarding publications and published gene lists were relationally linked in a MySQL database.

Results

Currently, 273 publications are inserted in Publime, containing 1282 gene lists. The analysis of MeSH terms allows classifying articles according to the classes of pathologies studied in the publication. Most represented categories of neoplasms are Breast, Prostatic, Ovarian and Colonic. A web interface allows biologists to retrieve information according to particular keywords. Information concerning a particular gene, a particular class of genes, or a particular class of studies can be easily queried and interpreted. Moreover a user can test for the existence of a significant overlap between a set of genes of interest and lists stored in Publime. The significance of the overlap can be estimated according to the hypergeometric distribution or Fisher's exact test. This approach helped us to identify common transcriptional targets of pRB, p16 and EWS/FLI (Ewing sarcoma breakpoint region 1) pathways having a different role in cell cycle control [2]. We annotated a total of 31243 human and 7476 mouse identifiers, representing 8187 human and 3991 unique genes, respectively. Interestingly, we observed that some genes are reported as differentially regulated much more frequently than others. Gene Ontology analysis revealed that the most frequently reported genes are predominantly involved in the regulation of cell cycle, regulation of cellular proliferation and in cell growth and mantainance.

Contact email: giacomo.finocchiaro@ifom-ieo-campus.it

References

1. Guffanti A, Finocchiaro G, Reid JF, Luzi L, Alcalay M, Confalonieri S, Lassandro L, Muller H: Automated DNA chip annotation tables at IFOM: the importance of synchronisation and cross-referencing of sequence databases. Appl Bioinformatics 2003, 2(4):245-249.

2. Finocchiaro G, Mancuso F, Muller H: Mining published lists of cancer related microarray experiments: Identification of a gene expression signature having a critical role in cell-cycle control. BMC Bioinformatics 2005, 6(Suppl 4):S14.

Intragenic antisense transcription correlates with long UTRs

Finocchiaro G (1,2), Parise P (1,2), Di Ninni V (1,2), Francois S (1,2),
Carro MS (1,2), Muller H (1,2)

(1) FIRC Institute of Molecular Oncology Foundation (IFOM), Milano
(2) European Institute of Oncology (IEO), Milano

Motivation
Studies on the distribution of transcription factor binding sites (TFBS) along entire chromosomes by chromatin immunoprecipitation have revealed that the majority of binding sites for several transcription factors lie far away from annotated promoters. Evidence is accumulating that the amount of transcribed DNA may be much higher than previously thought, with a significant fraction of transcription occurring in the antisense direction of annotated genes. Deciphering the regulatory potential of antisense transcripts is still in its infancy.

Methods
We performed genomic mapping of the 5'-ends of antisense transcripts to corresponding sense transcripts with the aim of identifying hotspots of intragenic antisense transcription. Correct orientation of antisense transcript was evaluated using criteria simlar to those adopted by [1]. When multiple antisense transcripts were mapped close to each other, they were assigned to distinct Antisense Transcription Starting Regions (ATSR) if their 5'ends were more than 500 bp apart. Otherwise they were classified as being part of the same ATSR. UTR length was estimated based on annotations available at
http://hgdownload.cse.ucsc.edu/goldenPath/hg17/database/refSeqSummary.txt.gz, DBTSS (http://dbtss.hgc.jp/), and a recently published curated dataset of 5'UTRs [2], with highly consistent results.

Results
A total of 7903 ATSRs was identified that were mapped onto 5075 genes. In agreement with published results, we find that antisense transcripts are particularly abundant at the 5' end as well as at the 3' end of genes. However, our map identifies the first exon, the 5'end of the first intron, and the 5'end of the last exon as hotspots of intragenic antisense transcription when the number of antisense transcripts in a given region is normalized per unit sequence. In particular, we identified 654 genes with an ATSR in their first exon, 1847 genes with an ATSR in their first intron, and 756 genes with an ATSR in their last exon. The remaining ATSRs are distributed evenly along loci. Our findings are supported by the enrichment of known transcription factor binding sites in the vicinity of ATSRs as well as by binding of the general transcription factor TAF1 as measured by chromatin immunoprecipitation on genomic tiling arrays [3]. The vicinity of hotspots of antisense transcription to the UTRs of sense transcripts prompted us to explore the UTRs of genes with evidence of intragenic antisense transcription in more detail. We find that the presence of antisense transcripts is strongly and positively correlated with the length of UTRs of the corresponding sense transcript. While the median length of 5'UTRs of genes in the genome was found to be 142 bp, genes with an ATSR in their first exon had a median 5'UTR of 241 bp. A T-test performed on the log transformed values of 5'UTR lengths indicates that this is significant at P = 7.67E-24. Similarly, while the median length of 3'UTRs in the genome was estimated to be 636 bp, genes with an ATSR in their last exon had a 3'UTR of median length 1225 bp. T-test on log transformed 3'UTR lengths indicates that this result is significant at P = 1.63E-29. Potential implications of this finding for the regulatory role of antisense transcripts will be discussed.

Contact email: heiko.muller@ifom-ieo-campus.it

References

1. Chen J, Sun M, Kent WJ, Huang X, Xie H, Wang W, Zhou G, Shi RZ, Rowley JD: Over 20% of human transcripts might form sense-antisense pairs. Nucleic Acids Res 2004, 32(16):4812-4820.

2. Davuluri RV, Suzuki Y, Sugano S, Zhang MQ: CART classification of Human 5'UTR Sequences. Genome Research 2006, 10:1807-1816.

3. Kim TH, Barrera LO, Zheng M, Qu C, Singer MA, Richmond TA, Wu Y, Green RD, Ren B: A high-resolution map of active promoters in the human genome. Nature 2005, 436(7052):876-880.

A statistical empirical energy function for proteins

Fogolari F (1), Pieri L (1), Bortolussi L (2), Dovier A (2)

(1) Dipartimento di Scienze e Tecnologie Biomediche University of Udine Piazzale Kolbe, 4 33100 Udine - Italy
(2) Dipartimento di Matematica e Informatica University of Udine Via delle Scienze 206 33100 Udine Italy

Motivation
Reduced representations of proteins have been playing a keyrole in the study of protein folding. Many such models are available, with different details of representation. The aim of the present work is to provide both a discrete and an analytical forcefield for a reduced model employing only two centers of interactions per amino acid.

Methods
All protein structures in the set top500H have been converted in reduced form. The distribution of pseudobonds, pseudoangle, pseudodihedrals and distances of centers of interactions have been obtained and fitted to analytical functions. The fit allows to link the features of the distributions to specific secondary structure elements. The correlation between adjacent pseudodihedrals has been converted in an additional energetic term which is able to account for cooperative effects in secondary structure elements.

Results
Tests have been performed using a minimization tool developed within a novel concurrent multi-agent framework for protein structure prediction. Tests will be presented also on decoys and on simple model systems. The pseudodihedral correlation term appears to be of utmost importance for proper discrimination of native like structures.

Contact email: ffogolari@mail.dstb.uniud.it

On the origin and evolution of biosynthetic pathways: integrating microarray data with gene structure and organization

Fondi M, Brilli M, Fani R

Dipartimento di Biologia Animale e Genetica dell'Università di Firenze, via Romana 17/19, 50125 Firenze, Italia

Motivation

The availability of the nucleotide sequence of complete genomes from an increasing number of organisms belonging to the three cell domains, Archaea, Bacteria and Eucarya is providing an enormous body of data concerning the structure and the organization of genes and genomes. As a result it is now possible to shed some light on the mechanisms involved in their evolution and responsible for the shaping of metabolic pathways. The emergence of basic biosynthetic pathways represent one of the major and crucial events during the early evolution of life. Their appearance and refinement allowed primitive organisms to become increasingly less dependent on exogenous sources of amino acids, purines, and other compounds. Among the theories proposed to explain how metabolic pathways have assembled and evolved, the patchwork hypothesis (Jensen 1976) is the most accepted one. Schematically it predicts that the extant pathways have been assembled starting from a restricted core of ancient genes that underwent gene duplication events (generating paralogous genes) followed by evolutionary divergence. By this "two-step" mechanism, ancient genomes may have increased their dimensions and/or gained novel metabolic abilities. In some cases, the divergence between paralogous genes (or set of genes) might be due to divergence in the regulation mechanisms controlling their expression and\or the regulation of enzymatic activity rather than mutations affecting catalytic sites or sites involved in the bindingof specific ligands. This is the case of genes involved in the so-called Common Pathway (CP) of lysine, threonine and methionine. These three biosynthetic routes share the first two steps, i.e. the phosphorylation of an aspartate molecule and the subsequent oxidation that lead to the formation of aspartyl phosphate. In the γ-proteobacterium Escherichia coli three different aspartokinases (AKI, AKII, AKIII, the products of thrA, metL and lysC, respectively) can perform the first step of the CP. Moreover, two of them (metL and thrA) are bifunctional, carrying also homoserine dehydrogenasic (hsd) activity, the final branching point of threonine and methionine routes. The second step of the CP is catalyzed by a single aspartate semialdeyde dehydrogenase (ASDH, the product of asd). Thus, in the CP of E. coli while a single copy of ASDH perform the same reaction for three different metabolic routes, three different AKs have evolved and been maintained to perfom a unique step. Why such a situation emerged and maintained? How is it correlated to the different regulatory mechanisms acting on these genes? The integration of data concerning gene structure, organization, and phylogenetic distribution, with information coming from the analysis of microarray experimental data represents a very powerful tool to elucidate the mechanisms and forces driving the assembly and the evolution of metabolic pathways. For this reason, we used such an approach to answer those questions mentioned above, focusing the attention on proteobacteria whose genome was fully sequenced.

Methods

Microarray data were downloaded as supplemental material to published papers. Database search was performed using BLAST software (Altschul et al. 1997). Phylogenetic trees were constructed using MEGA3 software (Kumar et al.2004). Microarray experiments data statistic analysis were conducted using R software (ver 2.1.1 R Development Core Team 2005).

Results

Structure and phylogenetic distribution of AK and HD coding genes in proteobacteria

Data obtained revealed that the presence of multiple copies of the AK coding gene and their fusion with HD domains are restricted to the Y-subdivision of proteobacteria. A model explaining the origin and evolution of AK and HD coding genes was depicted, which was also supported by a phylogenetic analysis of both AK and HD sequences. According to this model, the genome of the

proteobacterial ancestor harboured a monofunctional copy of both ask and hsd genes. Then a paralogous duplication of these genes concomitant to the fusion of the copies occurred within the γ-proteobacteria giving raise to a bifunctional gene that, in turn, duplicated generating the ancestors of the extant metL and thrA.

Organization of AK and HD coding genes in proteobacteria

The analysis of the organization of lysine, threonine, and methionine biosynthetic genes revealed that the appearance of fused genes was paralleled to the assembly of operons of different sizes, suggesting a strong correlation between the structure and organization of these genes.

Analysis on microarray data

In order to check the existence of a correlation between the structure, the organization and the coexpession of genes belonging to the same metabolic route, a statistic analysis of microarray data retrieved from experiments conducted on E.coli and Pseudomonas aeruginosa, showing different gene structure and organization, was carried out and data obtained will be discussed.

Contact email: r_fani@dbag.unifi.it

Supplementary informations

Microarray data references are available and not cited for shortness (10 publications, 79 conditions)

GORetriever: a novel Gene Ontology annotation tool based on semantic similaritiy for knowledge discovery in database

Fontana P (1), De Mattè L (1), Cestaro A (1), Segala C (1), Velasco R (1), Toppo S (2)

(1) IASMA Istituto Agrario di S. Michele all'Adidge (trento)
(2) Dipartimento di Chimica Biologica Università degli Studi di Padova

Motivation

Over the years, biological databases have grown at a spasmodic rate and we have assisted at an exponential increase of the available amount of data. The biological knowledge, associated to a particular sequence, is usually expressed in natural language and stored as free text. Most of the information is unstructured and does not follow strict semantic rules. End users can easily understand this human readable format but the same knowledge cannot be managed and caught by a computer program. The Gene Ontology (GO) (1) effort goes in this direction providing a structured vocabulary where each term is described as a father-child relationship and multiple inheritances are allowed. In this framework protein functions are represented by a DAG (Directed Acyclic Graph) starting from the root, consisting of general terms, to the leafs containing different levels of detailed descriptions. Such an ordered infrastructure makes feasible to infer and measure semantic similarities of distant or different concepts simply looking at the information content they share.

Methods

Our method is an approach to automatically annotate sequences based on retrieved GO terms. The starting list of GO terms to evaluate may be obtained, for instance, by a simple similarity search of the query sequence against a database of GO annotated proteins. The GO hits are processed in order to reconstruct all of the possible paths that lead to the root node. During the recursive process each node is scored adding the weights of the nodes encountered during the path reconstruction. As a result we obtain a trimmed GO graph consisting only of the terms found in the database search: for each term we keep track of its occurrence and of its cumulative score. The algorithm calculates the Z score of the cumulative score obtained for each node. The path, including the nodes with the highest weights, is extracted. Due to the additive property used to weight the nodes, only generic annotation terms are discriminated efficiently. These nodes are near the root node and therefore they are highly frequent. To solve this problem a different measure has been used to get a good tradeoff between detail information and statistical significance. The nodes belonging to the most probable selected starting path, may contain too many GO terms. They are, then, grouped on the basis of their Information Content (IC, based on the frequency of each term) and their semantic distance calculated applying the Lin (2) formula that quantify the amount of information shared. This clustering criterion of similar terms allows to restrict the searching space of correct annotations. The remaining term hits are then ranked efficiently using two statistical scores and an entropy based measure: "Internal Confidence" (InC), "Absolute Confidence" (AC) and Theil Index (TI). The InC and AC scoring methods have been specifically developed to assess the statistical significance of the retrieved hits and are both based on non-cumulative node weights divided by either cumulative root node weight (InC) or by the maximal theoretical weight (AC). Theil index (TI) (3) is derived from Shannon's measure of information entropy and it is applied to measure the inequality of score distribution over the trimmed GO graph. The final retrieval step is based on ranking GO terms with the highest score and information content (IC).

Results

GORetriever has been benchmarked using SwissProt Release 48.7. We randomly extracted from the database four sets of 5000 sequences sharing less than 10% of sequences one another. We tried to recover the original GO annotation using a simple BLAST search against SwissProt Release 48.7 from which test set sequences have been removed. Each GO term has been ranked according to its AC and InC score while TI index has been considered as an entropy measure of the GO distribution in the graph. Three different categories have been taken into account to evaluate GORetriever

performance: "exact" matches are GO terms identical to the original ones, "similar" matches are GO terms that are very close to the original ones (i.e father-child relationship), "mismatches" hits are the remaining cases. Almost the totality of the sequences in each test set have been annotated and the amount of correct hits that GORetriever has assigned is similar in each test set showing both high sensitivity and selectivity. On average 91% of the whole hits are "exact" matches, 3% are "similar" matches and 6% are "mismatches".

Contact email: paolo.fontana@iasma.it

References

1. Gene Ontology Consortium. The Gene Ontology (GO) project in 2006. Nucleic Acids Res. 1;34(Database issue):D322-6, 2006.
2. Dekang Lin. An Information-Theoretic definition of similarity. Proceedings of the Fifteenth International Conference on Machine Learning, pages 296-304, 1998.
3. Theil Henri. The measurement of inequality by component of income. Economics Letter, 2, 1979

HServ and SNPly: a software infrastructure and a web agent for linking information on genetic variation

Gallimbeni R, Falconi M, Desideri A

Department of Biology, University of Rome "Tor Vergata", Via della Ricerca Scientifica, 00133 Rome, Italy.

Motivation
The availability of public, complex and growing sources of data is a common trait of current research on human genetic variation. As a consequence, information increasingly needs to be located, filtered and pieced together.

Methods
HServ is a software framework for the quick development of web-agents in charge of collecting, gluing together and processing data from heterogeneous and possibly varying sources. Such agents should act both as a web server and a web client. They may react to user requests performing a variety of automatic remote searches, then extract appropriate contents, process filtered data, and eventually publish their responses on web pages. HServ is a collection of modules written in Haskell, a pure functional language. It is intended as a lightweight toolset for experimenting with different data-mining heuristics.

Results
SNPly, a web-agent devoted to the exploration of public databases and literature concerning SNPs, is being built on the basis of HServ. SNPly looks for semantic intersections in data from distinct sources, thus allowing researchers to detect implicitly related pieces of information which are not expressly linked to each other.

Availability: http://www.snply.org/

Contact email: mailto:rg@snply.org

Mitochondrial phylogeny of Anura (Amphibia): a case study of congruent phylogenetic reconstruction using amino acid and nucleotide characters

Gissi C (1), San Mauro D (2), Pesole G (1), Zardoya R (2)

(1) Dipartimento di Scienze Biomolecolari e Biotecnologie, Università di Milano, Via Celoria, 26 - 20133 Milano, Italy
(2) Departamento de Biodiversidad y Biología Evolutiva, Museo Nacional de Ciencias Naturales-CSIC, José Gutiérrez Abascal, 2; 28006 Madrid, Spain

Motivation

We explore whether phylogenetic analyses of the same sequence data set at the amino acid and nucleotide level are able to recover congruent topologies, as well as the advantages and limitations of both alternative approaches. As a case study, mitochondrial protein-coding genes were used to discern among competing hypotheses on the phylogenetic relationships of major anuran amphibian lineages.

Methods

To properly address this phylogenetic question, the complete nucleotide sequences of the mitochondrial genomes of two archaeobatrachian species, Ascaphus truei and Pelobates cultripes, were determined anew. Bayesian and maximum likelihood phylogenetic inferences of the same sequence data set were performed based on both amino acid and nucleotide characters, with the latter analysed either as codons or as a reduced data set of first+second (P12) codon positions. In addition, likelihood-based ratio tests were performed to evaluate the support of alternative topologies.

Results

The different data sets arrived at congruent and highly supported topologies, suggesting a similar phylogenetic resolving power of the two character types provided that correctly selected sites and appropriate evolutionary models are used. The reconstructed anuran mitochondrial phylogeny supports the paraphyly of Archaeobatrachia, with Ascaphus as sister group to all the remaining anurans, and Pelobates as sister group of Neobatrachia. However, the employed tree reconstruction methods and likelihood-based ratio tests seemed to be negatively affected by the fast evolving sequences of neobatrachians, suggesting that the phylogeny of Anura here presented is not definitive, and needs further investigation using extended taxon sampling.

Contact email: carmela.gissi@unimi.it

Supplementary informations

This work was supported by a BIODIBERIA project (European Commission - Human Potential programme) to C.G., a FIRB project (Ministero dell'Istruzione e Ricerca Scientifica, Italy), and a project of the Ministerio de Educación y Ciencia of Spain to R.Z. (CGL2004-00401)

Protein contact prediction with correlated mutation analysis using mixed physiochemical constraints

Horner D (1), Pesole G (2)

1) Dipartimento di Scienze Biomolecolari e Biotecnologie, Universita degli Studi di Milano
2) Dipartimento di Biochimica e Biologia Molecolare, Universita di Bari

Motivation

While models of sequence evolution used in many bioinformatic and evolutionary approaches assume that individual sites evolve independently of one-another, it has long been expected that in certain situations, for example amino acids that constitute structurally or functionally important contacts in proteins, should undergo a form of correlated, or compensatory evolution. This assumption forms the basis of many algorithms designed to predict protein structural contacts from alignments of homologous protein sequences. While such approaches have shown considerable potential, levels of accuracy typically fall short of those required to use correlated mutation analysis in structure prediction approaches. Interestingly, very few Correlated Mutation Analysis (CMA) algorithms incorporate phylogenetic information (the tree describing evolutionary relationships between the sequences under analysis). Furthermore, while several CMA algorithms incorporate models of biophysical properties of different amino acids, these typically assume that a single biophysical parameter governs all potentially correlated substitutions for a given pair of sites. We present here an updated version of our algorithm which incorporates phylogenetic information in the detection of pairs of sites undergoing potentially correlated evolution. In the current implementation, the biophysical parameters used to quantify the degree of correlation between pairs of sites are allowed to vary over the tree.Furthermore, we introduce an improved MonteCarlo data simulation procedure that allows rapid evaluation of the significance of results obtained.

Methods

Phyogenetic trees describing the relationships between homologs of proteins of known structure were estimated by standard methods. Ancestral sequences (at internal nodes on phylogenetic trees) were reconstructed under the maximum likelihood criterion. The magnitude of correlation between pairs of sites isevaluated by calculating correlation coefficients based on changes in biophysical characteristics implied by substitutions inferred on the relevant phylogenetic tree. The current implementation of this protocol allows the simultaneous calculation of correlation scores based on different biophysical parameters and the optimization of the correlation criterion in different parts of the tree. The significance of calculated correlation scores is evaluated using a MonteCarlo simulation in which a null distribution of correlation scores for every pair of sites is calculated by repeatedy independently distributing observed substitutions over the tree according to relative branch lengths.

Results

Allowing the biophysical parameter governing correlated substitutions between a pair of sitesto vary over the phylogenetic tree increases the sensitivity of CMA with our algorithm. The currently implemented measure of significance substantially decreases the number of false positive predictions with repect to a previous version which did not incorporate lengths of branches on the phylogenetic tree.

Contact email: david.horner@unimi.it

Improving the capacity of the CSTMiner algorithm to correctly classify conserved sequences

Horner D (1), Re M (1), Nasi C (1), Pesole G (2)

(1) Dipartimento di Scienze Biomolecolari e Biotecnologie, Universita degli Studi di Milano
(2) Dipartimento di Biochimica e Biologia Molecolare, Universita di Bari

Motivation
The CSTminer algorithm identifies sequences conserved between genomes (CSTs) through the use of a BLAST-like similarity search. A simple algorithm describing the evolutionary dynamics expected of codingsequences (a predominance of synonymous subsitutions and conservative amino acid changes) is used to ascribe a Coding Potential Score (CPS) to conserved elements. Such elements are classified as coding or non-coding through reference to results obtained from coding and non-coding "training sets".While in general this approach has proved extremely effective, the currently implemented methodology fails to unambiguously classify a significant number of (predominantly short) CSTs.

Methods
While the original method to evaluate the coding potential of CSTs relied on a simple threshold value derived from CPS values obtained empirically from known coding and non-coding sequences, we have developed a statistical measure of CPS scores derived from a simple data randomizationprocedure. Our method also incorporates observations of the "shadow effect" whereby a reverse reading frame tends, for coding csts, to exhibit a high coding potential. We thus calculate a P-value which indicates whether an observed CPS score can be explained by chance, given the composition, and level of conservation of a CST.

Results
Here we show that a simple bootstrap-like data randomization procedure can improve the accuracy of CST classification. Furthermore, we demonstrate that this approach is also effective in the classification of short conserved stretches, suggesting that it may also be of use in the detection of novel short exons.

Contact email: david.horner@unimi.it

ESTissue: A novel method to identify gene expression profile EST based

Iacono M (1), Mignone F (1,2), Anselmo A (1), Pesole G (3)

(1) Dipartimento di Scienze Biomolecolari e Biotecnologie - Università degli Studi di Milano
(2) Dipartimento di Chimica Strutturale e Stereochimica Inorganica - Università degli Studi di Milano
(3) Dipartimento di Biochimica e Biologia Molecolare - Università di Bari

Motivation

Genes are differentially expressed in different tissues, in different developmental stages or in pathological conditions. Differences in expression levels are often caused by diverse regulatory control mechanisms operating at the transcriptional, post-transcriptional and post-translational levels. The possibility of identifying clusters of genes that share the same expression profile (in terms of tissue specificity or temporal patterns of expression) could be of great significance in the investigation of underlying regulatory mechanisms. The availability of data generated by high-throughput approaches such as Expressed Sequence Tags (ESTs) allows the in silico analysis of gene expression in different tissues and physiological conditions. Several methodologies have been developed to characterize gene expression profiles from EST data. However, all such approaches have their own particular limitations, including both methodological and data access issues.

Methods

Public databases currently accommodate around 7.6 million human ESTs, of these, approximately 6.3 million of these are clustered in the Unigene database. However, Unigene clusters are assembled without consideration of the origin (tissue, developmental stage etc) of individual sequences. The presence of an EST in a tissue-specific library implies that the given gene or isoform is expressed in that specific tissue (or condition), moreover, it is reasonable to assume that the level of expression of a gene should be correlated with the number of ESTs. We use the information available for each EST to compare the expression pattern of genes in different tissues and we are thus able to identify both tissue specific genes and housekeeping (ubiquitously expressed) genes.

Results

We have developed a system (available through a web interface) which allows the user to identify and download sets of coexpressed genes. Selection can be performed on tissue specificity, developemental stage or pathological state. An automated evaluation of expression profiles is also preformed on user submitted genes.

Contact email: mailto:graziano.pesole@unimi.it

Reshaping the mtDNA circle: new insights from four newly sequenced ascidian genomes

Iannelli F (1), Griggio F (1), Pesole G (2), Gissi C (1)

(1) Dipartimento di Scienze Biomolecolari e Biotecnologie, Università di Milano, Italy
(2) Dipartimento di Biochimica e Biologia Molecolare, Università di Bari, Italy

Motivation
The mitochondrial genome (mtDNA) of vertebrates evolves following few rules: compact structure, constant gene content, almost frozen gene order except for minor changes involving tRNA genes, one major non-coding region involved in genome replication and expression, and a strong compositional asymmetry (Saccone et al. 1999). Surprisingly, the mtDNA of basal chordates and vertebrate ancestors, Tunicata, seems to follow a completely different evolutionary trend, characterized by many gene rearrangements even in cogeneric species (Yokobori et al. 2003; Gissi et al. 2004) and accelerated evolutionary dynamics.

Methods
In order to further investigate the peculiarities of mtDNA evolution in tunicates, we amplified by long PCR and completely sequenced the mtDNA of four ascidians: two Stolidobranchia species, Microcosmus sulcatus (Pyuridae) and Styela plicata (Styelidae), and two cogeneric Phlebobranchia species, Phallusia mammillata and Phallusia fumigata (Ascidiidae). Gene rearrangements were carefully investigated.

Results
The analyses confirm previous observations of a high rate of gene rearrangement in these genomes. The two mtDNAs of the genus Phallusia have undergone even more gene rearrangements than the two Ciona species. Only three gene pairs retain a conserved order between the two Pyuridae, Microcosmus sulcatus and Halocynthia roretzi (Yokobori et al. 1999). Moreover the only gene block conserved in all previously available tunicate mtDNAs - the cox2/cob pair - is not conserved in the organism Styela plicata. This situation confirms the hypothesis that the only constrain to conserve this gene block is an overlap between the ORFs (Gissi and Pesole 2003). Furthermore, base compositional variability, shortness of rRNA genes, and absence of a main non-coding region were confirmed as common features of ascidian mitochondrial genomes and indicate that the evolutionary dynamics of ascidian mtDNA markedly diverge from those of vertebrates.

Contact email: fabio.iannelli@unimi.it
graziano.pesole@biologia.uniba.it
carmela.gissi@unimi.it

References
- Gissi C, Pesole G (2003) Transcript mapping and genome annotation of ascidian mtDNA using EST data. Genome Res. 13:2203-12.
- Gissi C, Iannelli F, Pesole G (2004) Complete mtDNA of Ciona intestinalis reveals extensive gene rearrangement and the presence of an atp8 and an extra trnM gene in ascidians. J. Mol. Evol. 58:376-389.
- Saccone C, De Giorgi C, Gissi C, Pesole G, Reyes A (1999) Evolutionary genomics in Metazoa: the mitochondrial DNA as model system. Gene 238:195-209.
- Yokobori S, Ueda T, Feldmaier-Fuchs G, Paabo S, Ueshima R, Kondow A, Nishikawa K, Watanabe K (1999) Complete DNA sequence of the mitochondrial genome of the ascidian Halocynthia roretzi (Chordata, Urochordata). Genetics 153:1851-62.

- Yokobori S, Watanabe Y, Oshima T (2003) Mitochondrial genome of Ciona savignyi (Urochordata, Ascidiacea, Enterogona): comparison of gene arrangement and tRNA genes with Halocynthia roretzi mitochondrial genome. J. Mol. Evol. 57:574-87.

Supplementary informations

Acknowledgements: This work was supported by the COFIN project "Molecular evolution of innate immunity and mitochondrial genome in Ascidiacea" (Ministero dell'Istruzione, dell'Università e della Ricerca).

Comparative Genomics of OXPHOS gene families

Lanave C, De Grassi A, Saccone C

Istituto di Tecnologie Biomediche, Sede di Bari, CNR, Bari, Italy

Motivation

The gene transfer from the primitive mitochondrion to the nucleus stopped in the 800 Myr of Metazoan evolution, freezing the mitochondrial shape and size (with few exceptions) and reducing the gene content to a small contribution. The evolution of the mitochondrial genome and its dynamics has been extensively studied, also by our research group[1-3], but mitochondrial biogenesis and function are complex processes still far from being completely understood. The complete protein complement of mitochondria has still to be identified with both experimental and computational approaches having been applied to predict its quantitative value. It is currently rated at about 600-750 different proteins in yeast[4] and 600-1300 proteins for the human organelle[5]. Consequently, it is evident that the number of nuclear encoded proteins for the mitochondrion, even if still uncertain, is more than one, probably two order of magnitude higher than the number of mitochondrially encoded proteins. In this context, it seems clear that both the evolution and function of the mitochondrion itself and of the whole cell must be subject to forces modulating the interaction between the two genomes. In order to investigate the evolutionary relationship between mitochondrial and nuclear genomes, we have focused our attention on nuclear gene families involved in the main mitochondrial function, i.e OXPHOS (Oxidative Phosphorylation). We have observed that OXPHOS nuclear genes have a lower trend to produce or preserve duplications in Metazoa [6] Here we report the phylogenetic analysis of some OXPHOS gene families: LBP carrier and Cytochrome c (Cytc), which are instead subject to a higher expansion trend. The integrated phylogenetic and expression analysis of these large OXPHOS gene families could lead to a better understanding of their high expansion trend and could be useful to investigate and correlate the expression level of gene family members with a tissue specific and/or developmentally specific role.

Methods

Sequence data All Metazoan members of the LBP gene family were retrieved using the human protein sequences P1, P2 and P3 as query for TBlastN search on cDNA databases of 16 Metazoa. Cytc protein sequences, already annotated in the SWISSPROT and NCBI database, were collected and used as TBlastN queries to search for duplicated genes in the ENSEMBL database on cDNA databases of 19 Metazoa. All the source transcripts are shown in De Grassi et al. paper[7] Phylogenetic analysis Protein sequences were multialigned and the alignments of nucleotide sequences were deduced by over-imposing protein multi-alignment. The full alignment length is: 153aa for LBP sequences and 121aa for Cytc sequences.. The Bayesian analysis was carried out using the MrBayesv3.0b4 program, with the General-Time-Reversible substitution model[1] for nucleotide sequences and the Dayhoff model for protein sequences.

Results

A peculiar property of LBP is that it is mitochondrially encoded in yeast (e.g. S. cerevisiae and S. pombae), while it has been transferred to the nuclear genome both in plants and animals. In the human genome, LBP is represented by a gene family of three members encoding three different isoforms (P1, P2 and P3). Cytc is a central component of the electron transfer chain and is involved in both aerobic and anaerobic respiration. It also takes part in other cellular processes such as apoptosis and heme biosynthesis. Both the LBP and Cytc gene families are exceptions to the general rule for which OXPHOS genes are under-duplicated in Metazoan genomes[7]. The phylogenetic analysis of the ATP synthase LBP suggests that (1) the three isoforms were already present before the Birds-Mammals divergence, (2) only the P3 isoform possesses a putative orthologous gene in all analysed Vertebrates and (3) the P1 isoform is the most evolutionary divergent. In addition, in silico analysis has shown that: (4) P1 has a lower expression level than P2

and P3 isoforms both in man and mouse and (5) P1 is not in the NRF1 regulation circuit. Further studies will be required to define the specific role of the three LBP gene family members, but we observed that the P1 isoform possesses evolutionary and functionally divergent features. Our analysis of the Cytc gene family give important additional information to previous studies. On the whole, this study underlines that the evolutionary history of these two gene families has followed a completely different destiny in Mammals: (1) the LBP gene family is highly conserved, presenting three functional isoforms in all the analysed Mammals and P1 specific features, common to both human and mouse genomes; (2) in contrast, the Cytc gene family has been subject to complex genomic and functional events only in Mammals, leaving a unique somatic isoform in the human genome and, at least, two functional isoforms (somatic and testis-specific) in rodent genomes, which has been previously found in all OXPHOS duplicated genes in Insects[8]

Contact email: cecilia.lanave@ba.itb.cnr.it

References

1. Lanave C., Preparata G., Saccone C., G. Serio. A new method for calculating evolutionary substitution rates. (1984) J. Mol. Evol., 20:86-93.
2. Saccone, C., De Giorgi, C., Gissi, C., Pesole, G., A. Reyes, Evolutionary genomics in Metazoa: the mitochondrial DNA as a model system. (1999) Gene 238, 195-209. Review.
3. Saccone C, Gissi C, Lanave C, Larizza A, Pesole G, A. Reyes, Evolution of the mitochondrial genetic system: an overview. (2000) Gene. 261(1):153-159. Review
4. Prokisch H, Scharfe C, Camp DG II, Xiao W, David L, Andreoli C, et al. Integrative analysis of the mitochondrial proteome in yeast. (2004) PLoS Biol. 2(6):e0795-e0804.
5. Taylor SW, Fahy E, Zhang B, Glenn GM, Warnock DE, et al. Characterization of the human heart mitochondrial proteome. (2003) Nat Biotechnol. 2003 Mar;21(3):239-240.
6. De Grassi, A., Caggese, C., D'Elia, D., Lanave, C., Pesole, G., C. Saccone, Evolution of nuclearly encoded genes in Metazoa. (2005) Gene ;354:181-188.
7. De Grassi, A., Lanave, C. and C. Saccone. Evolution of ATP synthase subunit c and Cytochrome c gene families in Metazoa. (2006) Gene Feb 3, In press
8. Tripoli, G., D'Elia, D., Barsanti, P., C. Caggese, Comparison of the oxidative phosphorylation (OXPHOS) nuclear genes in the genomes of Drosophila melanogaster, Drosophila pseudoobscura and Anopheles gambiae. (2005) Genome Biol. 6(2),R11.

Selection and ranking of genes relevant for cancer diagnosis based on the classification ability of their expression pattern

Maglietta R (1), D'Addabbo A (1), Piepoli A (2), Perri F (2),
Liuni S (3), Pesole G (3,4), Ancona N (1)

(1) Istituto di Studi sui Sistemi Intelligenti per l'Automazione, CNR, Via Amendola 122/D-I, 70126 Bari, Italy
(2) Unità Operativa di Gastroenterologia, IRCCS, "Casa Sollievo della Sofferenza"-Ospedale, Viale Cappuccini, 71013 San Giovanni Rotondo (FG), Italy
(3) Istituto di Tecnologie Biomediche-Sezione di Bari,CNR, Via Amendola 122/D, 70126 Bari Italy
(4) Dipartimento di Biochimica e Biologia Molecolare - Università di Bari, Via E. Orabona 4, 70126 Bari, Italy

Motivation
One of the main problems in cancer diagnosis by using DNA microarray data is the selection of genes whose expression is most significantly altered by the pathology by analyzing their expression profiles in normal and tumour tissues. The question we pose is the following: how do we measure the relevance of a single gene in a given pathology?

Methods
A gene is relevant for a particular disease if we are able to correctly predict the occurrence of the pathology in new patients on the basis of its expression level only. In other words, a gene is informative for the disease if its expression levels are useful for training a classifier able to generalize, that is, able to correctly predict the status of new patients. In this paper we present a selection bias free, statistically well founded method for finding relevant genes on the basis of their classification ability.

Results
We applied the method on a colon cancer data set and produced a list of relevant genes, ranked on the basis of their prediction accuracy. We found that, among more than 6500 available genes, 54 genes over-expressed in normal tissue and 77 genes over-expressed in tumour tissue showed prediction accuracy greater than 70% with p-value $p \leq 0.05$. The relevance of the selected genes was assessed a) statistically, evaluating the p-value of the estimate prediction accuracy of each gene; b) biologically, confirming the involvement of many genes in generic carcinogenic processes and particularly for the colon; c) comparatively, verifying the presence of these genes in other studies on the same data-set.

Contact email: ancona@ba.issia.cnr.it

The Wnt Pathway as in silico disease model for neuro-oncology

Malusa F, Gonzalez Couto E, Rossi M. Bakker A, Kremer A, Terstappen GC

Discovery Research, Sienabiotech S.p.A., Siena, Italy

Motivation

Rational drug design requires a profound understanding of the events taking place at the biomolecular level in the cell. Signaling and metabolic pathways are thus becoming increasingly important for the drug discovery process. This in-silico analysis of gene and protein interaction networks can be helpful in many ways for example to bring together and structure the present status of scientific knowledge. Specifically the Wnt pathway is relevant within neuro-oncology diseases like gliomas and medulloblastomas as well as in Alzheimers disease. Wnt signaling pathway plays a key role in essential cancer relevant processes.

Methods

We created a Wnt pathway mainly by using a combination of tools and sources that allowed the creation of a pathway map; this map was built using the LION Bioscience Pathway editor which contains database derived annotations as well as manually inserted annotations such as literature references and literature summaries. The platform supports the direct linking to relevant information (UniProt, ENSEMBL, OMIM, PDB, MedLine etc.) and is furthermore linked to experimental data and results from in-house information coming from wet lab (quantitative RT-PCR experiments), 2D-gel, MS/MS and in-silico analysis. The problem of overlapping and connecting pathways was handled by incorporating links between them and embedding them into presented pathway.

Results

At SienaBiotech S.p.A. we constructed a Wnt pathway to support target identification as part of the drug discovery process. It has impacted projects covering brain cancer and Alzheimer's disease, delivering an overview of the proteins and available information. The pathway is used as a storage, visualization and communication tool to support the interdisciplinary work involving biologists, neurobiologists, oncologists and bioinformaticians. This leads to a better understanding of the disease mechanisms in CNS, and can help to unravel unknown mechanisms involved in the Wnt signaling pathway.

Contact email: fmalusa@sienabiotech.it

GALT-Prot database: a database of the structural features of GALT enzyme and its mutations

Marabotti A (1), Festa M (1,2), D'Acierno A (1), Facchiano A (1)

(1) Istituto di Scienze dell'Alimentazione, CNR, Avellino
(2) Facoltà di Ingegneria, Università degli Studi di Napoli "Federico II", Napoli

Motivation

GALT enzyme (Galactose-1-phosphate uridylyltransferase) is involved in the galactose metabolism by catalyzing the conversion of galactose-1-phosphate and uridine-5'-diphosphate-glucose into glucose-1-phosphate and uridine-5'-diphosphate-galactose. The genetic disorder called "classical galactosemia" or "galactosemia I" (OMIM: 230400) is linked to the impairment of this enzyme. This disease can be potentially lethal if not detected early and it is revealed through symptoms such as gastrointestinal complaints, hepatomegaly, cataracts, mental retardation, and ovarian failure in females. These last two dysfunctions can persist even with a life-long dietary restriction. Classical galactosemia is characterized by a high allelic heterogeneity, and to date more than 150 different base changes were recorded in several populations and ethnic groups. The three-dimensional structure of the human enzyme has been created by homology modelling methods [1]. On the basis of this model, it is now possible to investigate the position and the influence of each single mutation on the structure and on the dimeric assembly of the enzyme, with the aim of explaining molecular events related to this pathology.

Methods

The 3D model of GALT is available in PDB data base (PDB code: 1R3A). Protein structure analysis software as DSSP, HBPLUS, NACCESS have been used to evaluate secondary structure, H-bond formation, solvent accessibility. Information about gene mutations have been found in literature [2] and in the public data base of GALT mutations at genetic level (GALTdb) [3]. To create the database, the Entity-Relationship diagram has been structured and then translated into a relational model; the database has been then realized using the well know open source RDBMS PostegreSQL [4], paying attention to indexes to be created to improve the whole performance. To realize the Web Application we used STRUTS [5], a well known framework that implements the Model 2 approach, a widely adopted variant of the Model-View-Controller design paradigm. Here a Controller servlet acts as a controller for the whole application while the business logic resides into java beans and other helper classes (the Model). The presentation layer (the View) has been clearly realized using JSP pages and tag libraries.

Results

To model the Web Application we have first of all identified two typical users of the system: the Administrator and the Public User. The Public User, for which login is not required, browses data and has the opportunity to submit new mutations to the Administrator. The Administrator, among other things, inserts, updates and deletes mutations; for the Administrator a login is clearly required. For each typical user, and according to the UML specification [6], we have detailed the Use Cases Diagram. For each Use Case we have also defined the corresponding Sequence Diagram. The data have been modeled using an Entity-Relationship diagram where some entities are worth to be noted. The Protein entity, for example, has been introduced to make to final database capable of storing data not just for the GALT protein; the Chain entity, a weak entity whose occurrences are identified by a number and by the corresponding protein, is used to model amino acids chains. The Analysis entity is again a weak entity, identified by the element of the chain under study, that specializes into several entities (H-Bonds, DSSP, Monomer, etc). The occurrence of the Mutation entity represents a mutation to be stored; for each mutation we also consider an attribute Reference that stores the title of the paper describing such a mutation. The GALT-prot database reports information by the literature about known nucleotide mutations and the related amino acid mutation. The structural

features related to this amino acid are visualized with the aim of giving a possible explanation of the effect of the mutation, in terms of structure/function relationships. These information would be useful for all people involved in the biochemical study of this protein.

References

1. Marabotti A, Facchiano AM. Homology modeling studies on human galactose-1-phosphate uridylyltransferase and on its galactosemia-related mutant Q188R provide an explanation of molecular effects of the mutation on homo- and heterodimers. J Med Chem. 2005; 48: 773-779.
2. Elsas, LJ 2nd, Lai K. The molecular biology of galactosemia. Genet Med. 1998; 1: 40-48
3. GALTdb: http://www.alspac.bris.ac.uk/galtdb/
4. POSTGRES: http://www.postgresql.org
5. STRUTS: http://struts.apache.org
6. UML specification: http://www.uml.org

Availability: http://bioinformatica.isa.cnr.it/GALT

Contact email: amarabotti@isa.cnr.it

Molecular dynamics and docking simulation of the G216D mutation in the catalytic domain of activated protein C

Marino F (1), Morra G (1), D'Ursi P (2), Salvi E (3), Milanesi L (1), Faioni E (4), Rovida E (1)

(1) Institute of Biomedical Technologies, National Research Council, Segrate (Mi), Italy.
(2) Department of Science and Biomedical Technologies, University of Milano, Italy
(3) Department of Environmental Sciences, University Bicocca of Milano, Italy
(4) Hematology and Thrombosis Unit, Ospedale San Paolo, University of Milano, Italy

Motivation

Activated Protein C (PC) is an anticoagulant trypsin-like serine protease whose deficiency is associated with an increased risk for venous thrombosis. We present here a computationally based analysis of a new variant, G216D, identified in heterozygous patients with normal concentration and reduced activity of the protein. G216 is located on a surface loop in the vicinity of the catalytic S195, H57, D102 triad: it is highly conserved in serine-proteases family and has an important role in determining the specificity of substrate recognition. Ligand recognition in trypsin-like serine proteases is driven by the insertion of a positively charged side-chain into a specificity pocket. The ligand residue establishes electrostatic interactions with an aspartic in position 189 while two highly conserved glycines (216 and 226) shape the binding cavity to accomodate the ligand. Multiple alignment of trypsin-like family members indicated that the loop including G216 is strictly conserved. However it was interesting to observe that the substitution G216D is present in alpha-tryptase, and not in its beta counterpart. The two tryptases share 90% of identical residues but a shows a reduced activity and a different substrate affinity. X-ray structure comparison shows that the main difference falls within the segment of loop 215-225, involved in substrate binding. In alfa, a kink of the loop, prevents the insertion of the ligand in the binding pocket. Based on the overall structure conservation in trypsin-like family members and the high similarity of binding pockets of PC with beta-tryptase, we have simulated by molecular dynamics the structural effects of G216D substitution in PC. We also tested, by docking approach, the ability of the mutant to correctly bind the substrate. The reduced functionality of PC variant is interpreted in comparison with tryptase model.

Methods

The molecular model of the G216D variant was prepared from the heavy chain of the X-ray PC structure (pdb entry: 1AUT) by residue substitution. Molecular dynamics (MD) simulations were performed for native and variant PC in the canonical ensemble (NPT) using the program CHARMM with the all-atom force field CHARMM22. The proteins were solvated in a tetrahedral box of about 25000 water molecules. The system was relaxed with molecular mechanics and then MD calculations were started. The system was heated to 300°K for 25000 steps, equilibrated for 50000 steps and let to evolve for 6.5 ns with a time-step of 2ft. Docking analyses were carried out on final conformations from MD and on X-ray structures of alpha and beta-tryptase (1LTO and 1A0L respectively) with the corresponding co-crystallized ligands (PPACK and APA). Docking was performed with Autodock 3.05 using the Lamarkian Genetic Algorithm with a grid of 40X40X40 points, a spacing of 0.375Å and centered on the binding site.

Results

The objective of the MD was to explore the effect of mutation on the conformation of functional loop 215-225. Analysis of RMS deviation from the initial structures showed a local conformational change in the mutant associated with the distortion of the loop which is not observed in wt protein. The distorsion is responsible for the displacement of the D216 side chain compared to starting modelled structure and a separation of aspartic C-alpha of ~3Å. This effect is similar to that observed in alpha vs. beta-tryptase where a Ca's separation of ~4Å was found. As for alpha-tryptase, whose X-ray structure was obtained in absence of ligand (Marquardt Uet al. J.Mol.Biol.

2002, 321:419-502), the altered conformation of the loop 215-225 in PC variant, seems to be incompatible with a correct insertion and presentation of the ligand. To test this hypothesis, we have simulated by docking approach the binding of the PPACK inhibitor (co-crystallized with PC) to the mutant model. Docking results showed that the ligand is localized at the external surface of the protein probably due to the shape and electrostatic modification of the substrate binding pocket. As a consequence the ligand arginine cannot reach and establish favorable interactions with aspartic 189 localized at the base of the pocket as it seems to be deviated by the negative and the bulky side chain of D 216. We conclude that, in the G216D mutant, ligand interaction, if any, does not produce a correct orientation in the binding pocket and prevent the peptide proteolysis. This observation is in agreement with the reduced functionality observed in heterozygous patients carrying this mutation. It also confirms the key role of the loop 215-225 and highlights the importance of glycine in this specific position for serine-proteases function.

Contact email: ermanna.rovida@itb.cnr.it

Supplementary informations

This work was supported by European Project BioinfoGRID (Bioinformatics Application for Life Science) and will be available under the Italian MIUR-FIRB project LITBIO.

Logistic regression of controlled functional annotations of classified genes

Masseroli M, Bellistri E, Pinciroli F

Laboratorio di Informatica BioMedica, Dipartimento di Bioingegneria, Politecnico di Milano piazza Leonardo da Vinci 32, 20133 Milano, Italy

Motivation

When the value of a dichotomous variable, for example indicating presence or absence of a characteristic in a subject, depends on one or more other continuous or discrete variables, logistic regression can be used to predict the proportion of individuals who have the characteristics, or to estimate the probability that an individual will have the characteristic. In medicine logistic regression is generally used to estimate the probability that an individual, who has some characteristics or symptoms that influence the onset of a disease, will show such disease. In this case it is also used to calculate which of such characteristics or symptoms influence more significantly, and to which extent, the onset of the disease. In biomolecular medicine the same approach could be used to identify which functional characteristics better explain the binary classification of a set of genes, for instance obtained through statistical and clustering analyses of gene expression results from microarray experiments. In order to test effectiveness of this approach, we implemented a software module that allows performing logistic regression analysis of functional signature annotations of classified gene protein products.

Methods

As source of updated functional information we used the GFINDer genomic knowledge base. GFINDer (http://www.bioinformatics.polimi.it/GFINDer/) is a Web system we previously developed to gather controlled functional and phenotypic genomic annotations sparsely available in numerous different databanks accessible via Internet, and perform their comprehensive statistical analysis and mining. GFINDer is implemented in a three-tier architecture based on a multi-database structure that constitutes its genomic knowledge base. In the first tier, the data tier, a MySQL DBMS manages the knowledge base, which is kept updated by automatic procedures that automatically retrieve gene and protein annotations from several on-line public databanks as soon as new releases of them become available.

For the logistic regression we considered the following usual non linear equation:

$$\ln [p / (1 - p)] = b0 + b1\ x1 + b2\ x2 + \ldots + bi\ xi + \ldots + bn\ xn,$$

where p is the proportion of considered classified genes in two evaluated classes; xi are the proportions of considered genes in the two evaluated classes that present the i characteristics; and bi are the regression coefficients for the i characteristics. The absolute value of each of these bi coefficients indicates the importance of the corresponding i characteristic in contributing to the considered gene classification.

In order to solve the non linear equation, within GFINDer processing tier we used a straightforward Active Server Page and Javascript implementation of a standard iterative method to minimize the Log Likelihood Function, which is defined as the sum of the logarithms of the predicted probabilities of belonging to the first of the two evaluated classes for those considered genes belonging to that class, and the logarithms of the predicted probabilities of belonging to the second of the two evaluated classes for those considered genes belonging to that second class. The Null Model was used as starting guess for the iterations, i.e. all bi coefficients are zero and the b0 intercept is the logarithm of the ratio of the number of considered genes belonging to the first of the two evaluated classes to the number of considered genes belonging to the second class. Minimization is by Newton's method, with an elimination algorithm to invert and solve the simultaneous equations. No special convergence-acceleration techniques were used.

To interact with the DBMS server containing the genomic knowledge base on the data tier, we used Microsoft ActiveX Data Object technology and Standard Query Language, whereas we employed Hyper Text Markup Language and Javascript to implement a Web graphic interface for the user tier,

which is composed of any client computer connected to the Web server on the processing tier through an Internet/intranet communication network.

Results

In GFINDer Web system we implemented a Logistic Regression module that exploits controlled functional information contained within the GFINDer genomic knowledge base to allow executing logistic regression analyses of functional signature annotations of protein products of user-uploaded classified gene lists. Initial results are promising and show that the implemented logistic regression analysis helps in identifying which protein functional characteristics better explain the considered classification of a set of genes. Thus, it could support better interpretation of gene classes defined through statistical and clustering analyses of gene expression results from microarray experiments, and it could contribute to unveil new of biological knowledge about the considered genes.

Availability: http://www.bioinformatics.polimi.it/GFINDer/

Contact email: masseroli@biomed.polimi.it

Genomic annotation and statistical analysis of protein families and domains for functional investigation of gene lists

Masseroli M, Franceschini A, Maffezzoli A, Pinciroli F

Laboratorio di Informatica BioMedica, Dipartimento di Bioingegneria, Politecnico di Milano piazza Leonardo da Vinci 32, 20133 Milano, Italy

Motivation

Protein families and domains constitute one of the most useful information to understand protein functions and to gain insight into interactions among their codifying genes. Comprehension of domain structure of proteins within completed genomes is also fundamental for better understanding the evolutionary forces and emerging functions shaping genomes. The increasing number of proteins for which domain-based annotation is available hence represents an important background for computational genome-wise analyses. To allow performing comprehensive evaluations of gene annotations sparsely available in numerous different databanks accessible via Internet, we previously developed GFINDer, a Web server that dynamically aggregates functional and phenotypic annotations of user uploaded gene lists and allows performing their statistical analysis and mining (http://www.bioinformatics.polimi.it/GFINDer/). Exploiting protein information present in Pfam and InterPro databanks, we developed and added in GFINDer new original modules specifically devoted to exploration and analysis of functional signatures of gene protein products. They allow annotating numerous user classified nucleotide sequence identifiers with controlled information on related protein families, domains, and functional sites, classifying them according to such protein annotation categories, and statistically analyzing the obtained classifications.

Methods

GFINDer Web system is implemented in a three-tier architecture based on a multi-database structure. In the first tier, the data tier, a MySQL DBMS manages all considered genomic annotations stored in different relational databases. In two of them, we structured protein information from Pfam and InterPro comprehensive and manually curated collections of protein families, domains and functional sites. To associate a protein characteristic with the codifying gene, we considered the protein accession numbers associated with a gene, as provided by Entrez Gene database. In order to efficiently exploit the hierarchical “parent/child” relationships that can exist between InterPro entries, relationships that describe a common ancestry between entries, we first precomputed their hierarchical trees and structured them in a GFINDer database table. Then, within GFINDer processing tier, in Javascript and Active Server Page scripts, we implemented protein functional signature categorical analyses based on controlled protein family, domain, and functional site categories. Created analysis procedures employ hypergeometric and binomial distribution tests and the Fisher’s exact test to assess statistical significance of the over and under representation of categorical protein annotations in a group of user classified genes.
To interact with the MySQL DBMS server on the data tier, we used Microsoft ActiveX Data Object technology and Standard Query Language, whereas we employed Hyper Text Markup Language and Javascript to implement a Web graphic interface for the user tier, which is composed of any client computer connected to the Web server on the processing tier through an Internet/intranet communication network.

Results

In Pfam databank version 19.0 we found 8,183 protein family domain entries, and in InterPro databank release 12.0 we found 12,542 entries (8,945 protein families, 3,289 protein domains and 308 functional sites including post translational modifications, repeats, active and binding sites). Out of these entries, 3,254 (2,486 protein families, 743 protein domains and 25 functional sites) were grouped in 837 hierarchical trees of parent/child relations (574 of protein families, 252 of protein domains and 11 of functional sites). Parent/child protein family trees had a maximum of 6

levels, with an average of 414 entries per level, whereas protein domain trees had a maximum of 5 levels, with an average of 149 entries per level.
GFINDer modules developed for the exploitation of such structured data provide Protein Family & Domain Annotation, Exploration, and Statistics analyses. The Exploration Protein Families & Domains module allows to easily and graphically understand either how many and which protein families, domains, and functional sites are associated with each considered gene, or how many of the selected genes refer to each protein family, domain, or functional site. When uploaded nucleotide sequence identifiers are subdivided in classes (e.g. from clustering analysis of microarray results), the Statistics Protein Families & Domains module allows estimating relevance of Pfam or InterPro controlled annotations for the uploaded genes by highlighting protein signatures significantly more represented within user-defined classes of genes.
Thus, new GFINDer modules allow performing genomic protein function analyses that well complement previously provided phenotypic and functional evaluations in supporting better interpretation of microarray experiment results and unveil new biological knowledge about the considered genes.

Availability: http://www.bioinformatics.polimi.it/GFINDer/

Contact email: masseroli@biomed.polimi.it

Extraction of recurrent motifs synthetic genomic sequences via dictionary-based compression

Menconi G (1), Dionisi F (2), Marangoni R (3)

(1) Department of Mathematics, University of Bologna, Bologna.
(2) ProteoGen Bio S.r.l., Pisa.
(3) Department of Informatics, University of Pisa, Pisa.

Motivation

Linguistic analysis of symbol sequences has a natural application to genomic sequence analysis. The large extent of biological databases paves the way for both a large-scale query of recurrent words in complete genomes and a selective search of important motifs strictly related to determined functional meaning or specific for a gene family. Words should be selected in order to be reliable and faithful as to linguistics as to biology. From a technical point of view, the methods of motif extraction should be sufficiently fast and computationally light also to allow an on-line fruition to be set.

Methods

The proposed method is based on the use of CASToRe, a dictionary-based compression algorithm of the Lempel-Ziv family. The algorithm selects a dictionary by exact matches and parses the input sequence in some variable-length recurrent words. The algorithm CASToRe is a very efficient complexity detector and it was already successfully used in genome clustering and coding sequence identification in Prokaryotic genomes. The input sequence is parsed in subwords belonging to the final dictionary relative to the sequence. A weight function is defined on the dictionary and a score is assigned to each word, based on its occurrence and length. Eventually, the words with highest score are collected in the so-called "set of interesting words": intuitively, they are the longest words occurring repeatedly along the input sequence. Preliminary applications of such approach on real eukaryotic genomic sequences have produced hard to interpret results, due to the high complexity of eukaryotic genomes. To bypass this problem, we focused our interest on synthetic sequences, generates by simulating specific symbol frequency distributions, and investigating the behaviour of the dictionary words on these artificial conditions.

Results

We generated several data sets of: periodic, Bernoullian and noise-perturbed periodic binary strings, in order to test the validity of the method. As a first conclusion, the score selects the more distinctive patterns within the sequence; for instance, the only interesting word in periodic sequences is the period pattern, while in coding genomic sequences the set of interesting words is mainly made of codons. Going into deeper details on the performed analysis, for each data set we have taken under consideration the distribution of word length and word occurrence within each sequence of the set. Moreover, we have studied the weight function against word length. For what concerns periodic sequences, we have also focused on the dependence of the average word score on the period length; in the case of the Bernoullian sequences, the same index was studied with respect to the characteristic probability p; furthermore, we concentrated on how the noise intensity affects the results on some periodic strings.

Contact email: marangon@di.unipi.it

Evaluation of protein models quality using neural networks. Application to the ETHE1 protein

Mereghetti P (1), Papaleo E (1), Fantucci P (1), Tiranti V (2), Zeviani M (2), Mineri R (2), De Gioia L (1)

(1) Department of Biotechnology and Bioscience, University of Milano-Bicocca, Milano
(2) Unit of Molecular Neurogenetics, Institute ``Carlo Besta'', Milano

Motivation

The development of reliable and accurate evaluation tools to check the quality of protein models is crucial for the improvement of useful prediction methods. Several energy functions and scoring algorithms for evaluating protein structures have been proposed and can be divided into different categories depending on the physical principles and on the structural features of the models considered in the evaluation. In the present contribution, we apply a method based on a neural network, to discriminate among correct and incorrect protein models of the protein encoded by the ETHE1 gene, which have been recently identified in mutated forms in patients affected by ethylmalonic encephalopathy.

Methods

A database of protein decoys with known three-dimensional structure was build combining models generated by different prediction methods. On each model we have computed several structural parameters, such as secondary structure content, solvent accessible surface, radius of gyration, stereochemical parameters and the backbone root mean square deviation between the model and the respective native structure, as measure of structural accuracy. The parameters dataset obtained was submitted to a principal component analysis to reduce redundacy and noise, the reduced dataset was then used to develop a neural network ables to describe the relationship between the principal components space and the model accuracy. The neural network is a three-layers feed-forward network with 11 sigmoid neurons in the first layer, 8 sigmoid neurons in the second layer and 1 linear neuron as output. Distinct neural networks were trained on different training-set partitions, obtaining an ensemble of 50 neural networks. As prediction results we consider the median over the 50 networks predictions. The models of “ETHE1” were generated starting form different sequence alignements, obtained with various fold recognition and homology modelling server.

Results

The comparison of different ETHE1 models and the evaluation of their reliability, as obtained by the procedure outlined in Methods, has allowed to disclose structure-function relationship that complement available experimental data.

Contact email: paolo.mereghetti@unimib.it

Analysis system for Protein Surface Recognition

Merelli I (1), D'Agostino D (2), Clematis A (2), Milanesi L (1)

(1) Institute of Biomedical Technologies, CNR, Milano (Italy)
(2) Institute for Applied Mathematics and Information Technology, CNR, Genova (Italy)

Motivation

The study of the protein-protein interaction is extremely complex and a method of theoretical analysis to reduce the possibility of interaction to a small dataset of protein would be extremely useful. Discarding the methods based on the analysis of the system from an energetic point of view, a typically engineering approach has been adopted for modelling the protein surface and analyzing its characteristics in order to define correlation between different macromolecules. The analysis of the superficial complementarities of two proteins is the first step to determine which proteins have the possibility to interact between them. The information contained in the morphology of a macromolecular surface is in fact the most important characterization for the definition of the protein-protein interaction. In case of complementariness the docking analysis is directed on the electrostatic characteristics and on the chemical-physical abilities of the proteins. The complementariness represents an important discriminating factor, so a precise model of the superficial characteristics of a protein complex is crucial for the definition of a valid docking.

Methods

The protein modelling always relies on its atomic coordinates, obtained by magnetic nuclear resonance analysis or by crystallography studies. By estimating the occupation volume of each atom it is possible to define the volumetric characteristics of the macromolecule. This three-dimensional grid which describes the protein is therefore analysed through a software which extracts the molecular surface establishing which points of the volume are inside the surface and which instead are outside. To accomplish this task, a high parallelizable algorithm, called Marching Cubes, has been successfully used. The output of this computation is a triangular mesh which describes through its topology the characteristics of the protein surface. But the analysis of the surfaces which has a crucial role to define structural similarities between different macromolecules is a problem that introduces remarkable difficulties. The comparison between different triangular meshes is very challenging because very similar surfaces can be described by different data structure. The basic idea of this work is to use a description method based on a group of images produced by the projection of small pieces of surface on a set of object-oriented coordinate system that rely on a cylindrical vertex-normal reference. The collection of all this images, in coordination with the information about their reference system, describes the whole superficial morphology of a protein. To reduce the mesh noise produced by different descriptions of the same protein surface a bilinear transformation has been chosen to define the images. The projection of each vertex of the surface patch around the reference system falls in a bin of the plane to which four counters are associated. These bins are adjourned in function of the point distance inside the bin. After the projection of all the chosen vertices, each counter will assume a certain value and the ensemble of all the counters represent the image.

Results

Through this local description of the protein surface it is possible to decompose the information about its topology in a set of bi-dimensional images (Fig 1). To enforce the discrimination of the functional sites, the description images can be extracted from each protein in correspondence of the amino acids that have a key role from a function point of view. The comparison of images from different proteins can be performed using both linear correlation or other image-processing algorithms. Nevertheless, this computation is be very time-consuming if a search all against all is used. For this reason a Grid approach will be evaluated in the frame of European Project BioinfoGRID (Bioinformatics Application for Life Science) and the Italian MIUR-FIRB project LITBIO (Laboratory for Interdisciplinary Technologies in Bioinformatics).

Contact email: ivan.merelli@itb.cnr.it

Supplementary informations
BIOINFOGRID web site is available at http://www.itb.cnr.it/bioinfogrid LITBIO web site is available at http://www.litbio.org

Fig.1 Comparison between two images of the L-Chain and of the H-Chain Thrombin.

A new strategy to identify novel genes and gene isoforms: whole genome comparison of human and mouse

Mignone F (1,2,*), Re M (1,*), Horner D (1), Pesole G (3)

(1) Dipartimento di Scienze Biomolecolari e Biotecnologie, Università degli studi di Milano
(2) Dipartimento di Chimica Strutturale e Stereochimica Inorganica, Università degli studi di Milano
(3) Dipartimento di Biochimica e Biologia Molecolare, Università di Bari
(*) These authors contributed equally to this work

Motivation

Despite consistent efforts to improve the annotation of the human genome, we are still far from either having a complete list of human genes or knowing the correct structure of already "annotated" genes. Moreover certain genes are characterized by a very low expression level which complicates the detection of their expression products. It is commonly accepted that one of the most reliable way to predict and identify novel protein-coding genes is the alignment of interspecific genomic sequences. This approach is constantly acquiring greater importance because of the increase in the rate of generation of complete (or near complete) genome sequences. Such an approach is also expected to generally improve the accuracy of gene annotation We present here an improvement of a methodology we previously developed for the identification of genome regions likely encoding protein-coding genes based on the detection of clusters of potentially coding conserved sequence tags (CSTs).

Methods

In a previous study, we presented a clustering method (benchmarked on human chromosomes 15, 21 and 22) able to highlight groups of coding CSTs characterized by a density comparable with that observed in annotated genes. CSTs are obtained using CSTminer, a tool that performs a cross-genome blast-like alignment and then calculates a Coding Potential Score (CPS) to discriminate between coding and non-coding sequences. We present here a new implementation of the clustering protocol with improved sensibility and selectivity. Moreover we have analyzed the entire collection of syntenic regions of H.sapiens and M.musculus genomes.

Results

A whole genome set of more than 130,000 coding Conserved Sequence Tags (CST) was obtained from our analysis generating more than 10,000 CST clusters. The accuracy of our observations has been assessed with respect to annotations contained in the latest human and mouse EnsEMBL release. Beside clusters of CSTs corresponding to genes annotated in both human and mouse genomes, we identified clusters corresponding to genes annotated in one genome only. We also identified CST clusters not corresponding to any annotated gene in either human and mouse genomes - potentially corresponding to unannotated genes. These clusters have been compared to expressed sequences databases to provide support to their genic nature.

Contact email: mailto: graziano.pesole@biologia.uniba.it

Developement of a data mining system for human cell cycle data analysis

Milanesi L (1), Alfieri R (2), Merelli I (1)

(1) Institute of Biomedical Technologies, CNR, Milano (Italy)
(2) Department of Biotechnology and Bioscience, University of Milano Bicocca, Milano (Italy)

Motivation

The cell cycle is a complex biological process that implies the interaction of a large number of genes. Disease studies on tumour proliferation and de-regulation of human cell cycle have to face with the problem of finding as quickly as possible information related to all the genes that are involved in this cellular process. This work aims to implement a new resource which collects useful information about the human cell cycle to support studies on genetic diseases related to this crucial biological process. Some resources that collect many biological pathways, such as cell cycle, are available for different organisms, but in the state of art there are no specific resources for human cell cycle data integration. The most important resources are Kegg Pathway Database (http://www.genome.ad.jp/kegg/pathway.html) and Reactome (http://www.reactome.org/). Kegg acts in a larger range because it is a collection of pathway maps for metabolic processes, genetic and environmental data such as signal transductions and human diseases. Reactome is a resource for human biological processes which relies on information about single reactions grouped into pathways. Another resource is Cyclonet (http://cyclonet.biouml.org/index.html), a database specifically focused on the regulation of eukaryotic cell cycle. It is less integrated with other biological databases and it is less user-friendly than others.

Methods

"HCCdb" the "Human Cell Cycle Database" is a resource which relies on data taken from Kegg and Reactome. In particular genes involved in the complete cell cycle pathway, in apoptosis pathway and in MAP kinase signalling pathway are taken from Kegg, while genes involved in mitotic and checkpoint pathways are taken from Reactome. To integrate data, we query many resources to collect information related to each gene and protein previously selected. The database infrastructure is designed to make possible an automated data integration: by using a set of Perl libraries it is possible to query a set of selected biological databases retrieving information about genes and proteins. Moreover, we created a database automatic updating system through a pipeline that queries public databases to integrate new data in our resource. The database administrator can access to a specific page where he can insert a gene name and perform the pipeline for data integration. As result it occurs an updating of all tables of the database: in this way the resource can maintained up to date. The main goal of this work is the integration of data related to each gene or protein. For this reason users can query the database contents both inserting the gene/protein name or using the IDs of public databases. The query results page is a complete report and users can browse data using direct links to the different biological database from which data are taken. Users can also query the database using key-words: the results is a list of genes related to the query. HCCDb data are stored in a relational database and a MySQL server is used for this purpose. HCCDb has a "snowflakes" schema, which present the important information about genes and proteins in the inside tables, while collects auxiliary data in the outside tables. The HCCdb database is accessible through a web interface made up of a set of HTML pages dynamically generated from PHP scripts.

Results

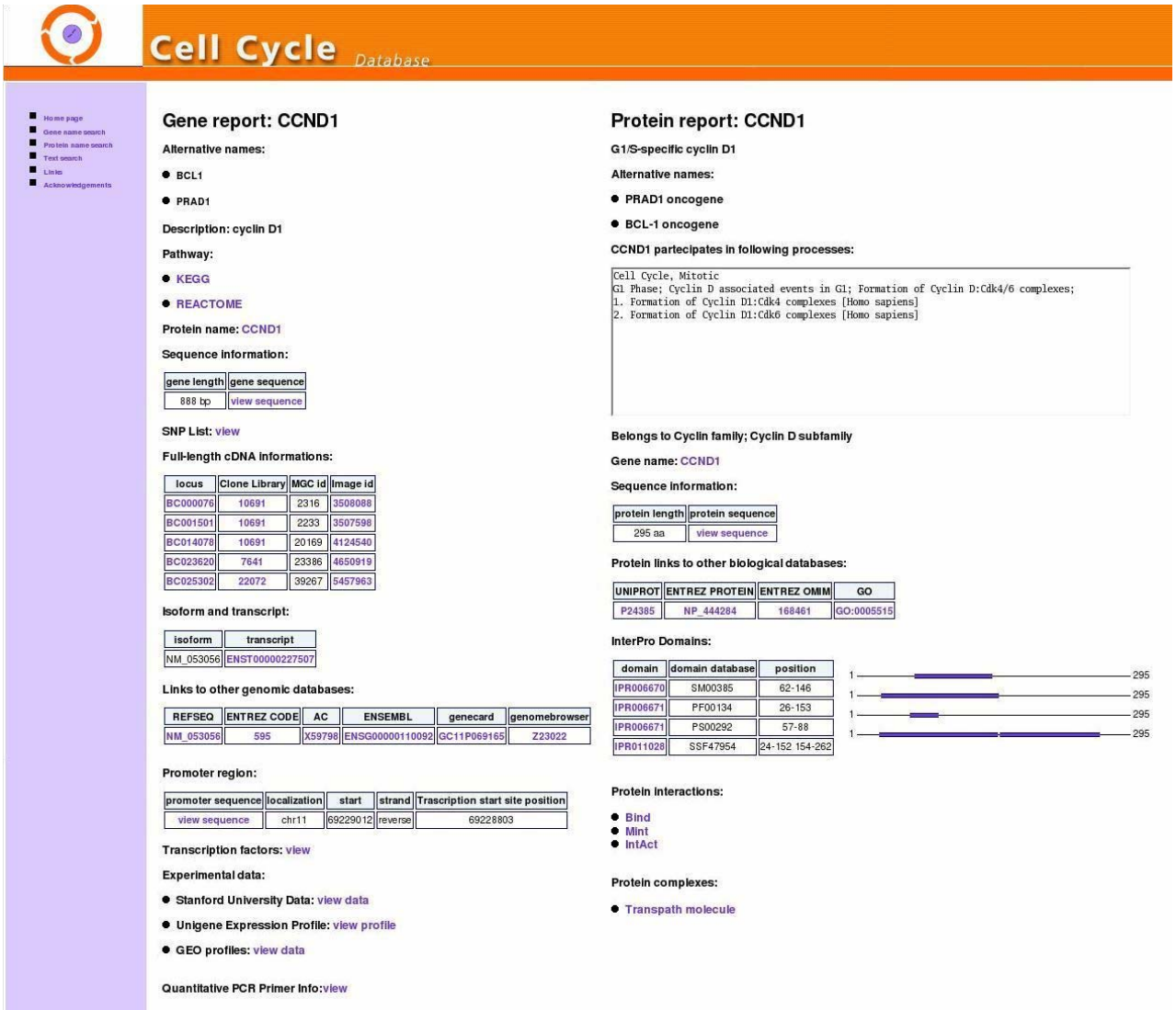
HCCdb is a resource that integrates as much as possible information related to genes and proteins involved in human cell cycle. The use of HCCdb has been tested while studying the Cyclin D1 genes, a regulator of the transition from G1 to S phase, which plays an important role in tumour-genesis. While investigating this gene, HCCdb has demonstrated its importance in retrieving information about experimental data, promoter and PCR primers that will be used to re-sequencing

this cell cycle regulator gene. This database has been realized in the frame of MIUR - LITBIO Project.

Availability: http://cellcycle.itb.cnr.it/

Contact email: luciano.milanesi@itb.cnr.it

Supplementary informations
The LITBIO web site is available at http://www.litbio.org



BioinfoGRID: Bioinformatics Grid Application for life science

Milanesi L (1), Andreas G (1), Arlandini C (4), Beltrame F (7), Bishop C (1), Breton V (3), Ernest P (5), Jacq N (3), Legre Y (3), Liò P (2), Liuni S (1), Mazzuccato M (6), Maggi G (6), Meloni G (4), Merelli I (1), Morra G (1), Orro A (1,4), Porro I (7), Sanger M (5), Shuai S (5), Trombetti G (1,4)

(1) CNR-ITB, National Research Council - Institute of Biomedical Technologies, Italy
(2) The Chancellor, Masters and Scholars of University of Cambridge, UK
(3) IN2P3-CNRS , Centre National de la Recherche Scientificque, Aubiere Cedex, France
(4) CILEA, Segrate, Italy
(5) DKFZ , Heidelberg, Germany.
(6) INFN, Italy
(7) DIST University of Genova, Italy

Motivation

The BioinfoGRID project is funded by the EU within the framework of the Sixth Framework Programme for Research and Technological Development FP6, as part of the specific programme 'Structuring the European Research Area', within the 'Research infrastructures' activity. 'Communication Network Development - eInfrastructure - Consolidating Initiatives'. The BioinfoGRID project web site will be available at http://www.itb.cnr.it/BioinfoGRID. The project aims to connect many European computer centres in order to carry out Bioinformatics research and to develop new applications in the sector using a network of services based on futuristic Grid networking technology that represents the natural evolution of the Web. More specifically the BioinfoGRID project will make research in the fields of Genomics, Proteomics, Transcriptomics and applications in Molecular Dynamics much easier, reducing data calculation times thanks to the distribution of the calculation at any one time on thousands of computers across Europe and the world by exploiting the potential of the Grid infrastructure created with the EGEE European project and coordinated by CERN in Geneva.

Methods

The BioinfoGRID projects proposes to combine the Bioinformatics services and applications for molecular biology users with the Grid Infrastructure created by EGEE (6th Framework Program). In the BioinfoGRID initiative we plan to evaluate genomics, transcriptomics, proteomics and molecular dynamics applications studies based on GRID technology. Genomics Applications in GRID - Analysis of the W3H task system for GRID. - GRID analysis of cDNA data. - GRID analysis of rule-based multiple alignments. Proteomics Applications in GRID - Pipeline analysis for domain search for protein functional domain analysis. - Surface proteins analysis in GRID platform. Transcriptomics and Phylogenetics Applications in GRID - Data analysis specific for microarray based on GRID servers. - To validate an infrastructure to perform Application of Phylogenetic based on execution application of Phylogenetic methods estimates trees. Database and Functional Genomics Applications - To offer the possibility to manage and access biological database by using the GRID EGEE. - To cluster gene products by their functionality as an alternative to the normally used comparison by sequence similarity. Molecular Dynamics Applications - To perform a challenge of the Wide In Silico Docking On Malaria (WISDOM project) - To improve the scalability of Molecular Dynamics simulations. - To perform simulation folding and aggregation of peptides and small proteins, to investigate structural properties of proteins and protein-DNA complexes and to study the effect of mutations in proteins of biomedical interest.

Results

BioinfoGRID will evaluate the Grid usability in wide variety of applications, the aim to build a strong and unite BIONFOGRID Community and explore and exploit common solutions. The BioinfoGRID collaboration will be able to establish a very large user group in Bioinformatics in EUROPE. This cooperation will be able to promote the Bioinformatics and GRID applications in EGEE and EGEEII. The aim of the BioinfoGRID project is to bridge the gap, letting people from

the bioinformatics and life science be aware of the power of Grid computing just trying to use it. The most natural and important spin off of the BioinfoGRID project will then be a strong dissemination action within the user's communities and across them. The BioinfoGRID project will provide the EGEEII with very useful inputs and feedbacks on the goodness and efficiency of the structure deployed and on the usefulness and effectiveness of the Grid services made available at the continental scale. In fact, having several bioinformatics scientific applications using these Grid services is a key moment to stress the generality of the services themselves.

Availability: http://www.itb.cnr.it/bioinfogrid

Contact email: luciano.milanesi@itb.cnr.it

Contribution to the ontology and system biology of muscle genes and application to microarray expression studies

Mittempergher L, Picelli S, Feltrin E, Colluto L, Nofrate V, Caldara F, Millino C, Campanaro S, Valle G

CRIBI Biotechnology Centre, Department of Biology, University of Padova, Padova

Motivation

Over the past ten years, our group has been working on the identification and characterization of skeletal muscle genes. A problem that remains open raises from the observation that vertebrate skeletal muscle consists of fibers having different contractile properties: fast and slow. Although several fast- and slow-specific genes have been known for many years, an overall picture of gene expression in these two types of muscle is still unclear. Therefore, we performed some gene expression experiments using microarrays, taking the mouse as a model organism because, differently from human, many of its muscles are mainly composed by a single type of fiber, making easier to delineate a "fast" or "slow" expression profile. A very effective approach to analyse microarray data is based on Gene Ontology (GO) (http://www.geneontology.org/), but there are two main problems: firstly, the list of probes needs to be continuously revised as the genome annotation and GO terms become better defined. Secondly, we found that the GO terms related to muscle were very scanty, making worthless their application to our studies. This gave us a good motivation to contribute to the GO project and to start a collaboration with the GO Consortium to enrich our domain of interest (neuro-muscular genes) with new terms. At the same time, the System Biology Markup Language (SBML, http://sbml.org/) can be used to annotate protein interactions.

Methods

Re-annotation of microarray probes. Oligo sets used for microarray experiments must be re-annotated because: 1) new genes and splice variants are continuously identified and frequently an oligo designed specifically for a gene cannot discriminate between splice variants or gene isoforms; 2) gene annotation in public databases is continuously updated. We have implemented a procedure that make use of GoMiner (http://discover.nci.nih.gov/gominer/) and other programs in order to obtain a faster annotation of our microarray platforms. GO and sub-ontology of neuro-muscular genes. The methodological approach is based on a deep understanding of specific problems by expert people. Every problem (for instance a group of proteins involved in a given process) is then "translated" into a series of GO terms that are submitted to the GO Consortium. In this respect we are collaborating with the GO curators at the EBI and we must acknowledge their help and availability. Once the new GO terms are defined, they can be associated to the corresponding genes. Since GO terms must be for general use, some very specific terminology cannot be included. Therefore we have started the development of a specific sub-ontology that could better explain some biological aspects such as muscle contraction or nervous system functions. This sub-ontology will complement the general GO terms. Annotation of protein interaction with SBML. GO terms are intended to describe proteins, not the network of their interaction and their functional regulation. Therefore, we annotate this information by SBML (level 2), using the CellDesigner software (Kitano et al. Nature Biotechnology 23, 961-966, 2005).

Results

Although this work of annotation is still in progress and will probably continue for some time, we have applied the methods described above for a whole-genome microarray expression analysis of mouse skeletal muscles using oligo microarray technology. We considered three muscles composed respectively of slow (soleus), fast (tibialis) and mixed fibers (gastrocnemius). Our preliminary results are shown in the table enclosed to the abstract, where the number of genes having an associated GO term rises from 6571 to 9933.

Contact email: lorenza.mittempergher@unipd.it

	old (OPERON) annotation	new (updated) annotation
Genes annotated with a GO term	**6571**	**9933**
Biological Process	**5430**	**8020**
muscle_contraction	**44**	**52**
muscle_development	**93**	**107**
heart_development	**65**	**74**
Cellular Component	**5575**	**8300**
cytoskeleton	**370**	**522**
actin_cytoskeleton	**91**	**135**
myosin (protein_complex)	**25**	**34**
Molecular Function	**5835**	**8739**
cytoskeletal_protein_binding	**147**	**215**
actin_binding	**115**	**159**
motor_activity	**61**	**88**

Environment specific substitution tables for thermophilic proteins

Mizugushi K (1,2), Sele M (3), Cubellis MV (3)

(1) Department of Biochemistry, University of Cambridge, UK
(2) Department of Applied Mathematics and Theoretical Physics, University of Cambridge, UK
(3) Dipartimento di biologia strutturale e funzionale, Universita' di Napoli "Federico II", IT

Motivation

Most organisms grow at temperatures from 20 to 50 °C, but some prokaryotes, including Archaea and Bacteria, are capable of withstanding higher temperatures, from 60 to >100 °C. Subtle differences between thermophilic and mesophilic molecules can be found when sequences or structures from homologous proteins are compared, but often they are family-specific and it is very difficult to derive general rules. The availability of complete genome sequences makes it feasible a large scale comparison between thermophilic and mesophilic proteins. Although most sequenced genomes of thermophilic organisms belong to archaea, a few are also available for eubacteria. We made independent comparisons of mesophilic proteins with their thermophilic counterparts of archaeal or eubacterial origins, since different mechanisms for the adaptation of proteins at high temperatures might have been exploited in the two kingdoms. Moreover we derived amino acid substitution tables that give the likely substitutions of amino acids in particular local environments because the conservation of amino acid residues has been shown to be strongly dependent on the environment in which they occur in the folded protein.

Methods

A database of 19168 protein sequences derived from the genomes of 10 archaea living at or above 60 °C was compiled. First, 3763 protein structures belonging to 1057 different families were taken from HOMSTRAD, a database of protein structure alignments for homologous families (1). The sequence corresponding to each structure was used as a query to search with BLAST the database of archaeal thermophilic proteins. In this way, we built alignments, where the first sequence is for a protein of known structure and the other ones are for its homologues from archaeal thermophiles. The residues of the first protein, whose structure is known, were assigned to eight different structural environments, i.e. alpha helix (buried or exposed), beta strand (buried or exposed), positive mainchain phi angle (buried or exposed) and coil (buried or exposed). Environment-specific amino acid substitution tables were calculated using the modified version of SUBST (K. Mizuguchi, unpublished). Substitution frequencies represent the likelihood of acceptance of a mutational event by a residue in the first sequence and in a particular structural environment, leading to any other residue in the archaeal thermophilic sequences. These tables, specific for thermophilic archaeal sequences, were compared to the standard substitution tables used in FUGUE (2). Environment specific substitution tables for thermophilic eubacteria were derived with the same method. In order to determine the effect of different substitution tables on sequence alignments, we used pairs of structures (one thermophilic and the other mesophilic) from HOMSTRAD. MELODY in the FUGUE suite of programs (2) was used to derive two profiles from the mesophilic protein structure; the first was obtained by using standard substitution tables and the second exploiting thermophile specific substitution tables. Each profile was aligned with FUGUE (2) to the sequence of the thermophilic protein, resulting in two alignments. The original structural alignment between the mesophilic and thermophilic proteins stored in HOMSTRAD, was used as a reference.

Results

A few general rules for the adaptation of proteins at high temperatures have been put forward so far. Our substitution tables derived from an extremely high number of raw substitution counts allowed us to confirm or disprove some of them. For instance it has been claimed that an increase in thermostability is correlated with the location of branch points in amino acids and beta and gamma branched amino acids increase protein thermostability. Our substitution tables derived for thermophilic archaea confirm that in most environments Ile is a preferred amino acid. We aligned

the sequences of thermophilic proteins to those of mesophilic proteins of known structure. These are the kind of alignments that must be used to model a thermophilic protein based on a mesophilic template. Unfortunately only a few pairs of homologous protein structures were available, where one member is mesophilic and the other thermophilic. On this limited set, we compared the alignments that exploited thermophile specific substitution tables and those with the standard substitution tables. The two kinds were comparably accurate, although for some families, thermophile specific substitution tables produced more accurate alignments.

Contact email: cubellis@unina.it

References

1. Mizuguchi K, Deane CM, Blundell TL, Overington JP. Protein Sci 1998;7:2469-2471
2. Shi J, Blundell TL, Mizuguchi K. J Mol Biol 2001;310:243-257

Molecular dynamics simulations of TBP complexed with diverse TATA variants

Morra G, Milanesi L

Institute of Biomedical Technologies, CNR, Segrate (MI)

Motivation

Promoters of many genes transcripted by RNA polII contain a consensus sequence TATAt/aAt/aX called TATA box, which is positioned 25-30 basepairs upstream of the transcriptional binding site. This sequence is recognized by the transcription factor TFIID, which binds to DNA via its subunit TBP. The crystallized complex TBP-DNA shows that TBP binds to the minor groove of the TATA element, which is strongly bent towards the major groove. Binding affinity and transcriptional efficiency critically depend on the DNA sequence. Recently, a set of diverse DNA patterns was analyzed by a neural network method, providing a ranking in terms of probability of being TATA sequences. This work aims at correlating the proposed classification of TATA sequences to structural features of the corresponding DNA-TBP complexes, as they emerge from molecular dynamics simulations.

Methods

The crystal structure of a human TATA-TBP complex (PDB code 1C9B) was used as starting model for four different low and high score DNA patterns. Mutations in the TATA sequence were modelled with CHARMM and the CHARMM27 force field starting from the wild type structure. After solvating the complex using the TIP3 water model, a system of orthorombic shape made of about 25000 atoms is considered for NPT MD simulations with CHARMM, timestep 2 fs and Verlet algorithm. Trajectories on the nanosecond scale are obtained for each DNA pattern. Structural analysis is carried out at two levels: first, the bending of DNA during dynamics is studied by means of MADBEND and CURVES. Second, the protein-DNA interface is analyzed in terms of hydrogen bonding pattern and van der Waals contacts.

Results

The analysis suggests that TATA sequences which obtain a high score in the neural network might be correlated to a high flexibility of the DNA segment when complexed with TBP, given by significant fluctuations of the conformation and larger bending. In order to interpret this result the interactions at the protein-DNA interface, as a way to identify sequence dependent differences, are under investigation.

Contact email: giulia.morra@itb.cnr.it

Alignment of Homologous Protein Structures in the Presence of Domain Motions

Mosca R (1), Schneider RT (1,2)

(1) FIRC Institute of Molecular Oncology Foundation, Via Adamello 16, 20139 Milan, Italy
(2) European Institute of Oncology, Via Ripamonti 435, 20141 Milan, Italy

Motivation

Structural alignment is an important step in protein comparison. Well-established methods for solving this problem in the case of rigid structures already exist. Methods for three-dimensional alignment in the presence of domain motions have only been developed in recent years (Ye and Godzik 2003; Shatsky, Nussinov et al. 2004).

Methods

Here, we present a new method for the flexible alignment of homologous protein. In this method the space of all possible alignments is represented as a graph on a set of potential matching fragments in the two molecules, scored on the base of their structural similarity. Every path in the graph represents a possible alignment while the optimal alignment is considered to be the longest path in the graph, i.e. the path yielding to the maximum score. A scoring strategy based on difference distance matrices is used. The alignment algorithm is coupled with a genetic algorithm for the identification of conformationally invariant parts (Schneider 2002) to define subsets of atoms suitable for group-wise least-squares superposition The alignment is characterized by the number of superimposed atoms and a measure, RMSDflex, that summarizes the RMSD for the superposition of several rigid regions.

Results

The alignment tool based on this method is able to align homologous structures with large hinge motions, such as two structures of the molecular chaperon GroEL from two different species. When compared to a method like DaliLite (Holm and Sander 1993) by aligning a set of homologous kinases structures with both methods, the new algorithm reaches good performance in terms of coverage and RMSDflex, while requiring only about one fifth of the CPU-time used by Dali.

Contact email: roberto.mosca@ifom-ieo-campus.it

References

- Holm, L. and C. Sander (1993). "Protein structure comparison by alignment of distance matrices." J Mol Biol 233(1): 123-38.
- Schneider, T. R. (2002). "A genetic algorithm for the identification of conformationally invariant regions in protein molecules." Acta Crystallogr D Biol Crystallogr 58(Pt 2): 195-208.
- Shatsky, M., R. Nussinov, et al. (2004). "FlexProt: alignment of flexible protein structures without a predefinition of hinge regions." J Comput Biol 11(1): 83-106.
- Ye, Y. and A. Godzik (2003). "Flexible structure alignment by chaining aligned fragment pairs allowing twists." Bioinformatics 19 Suppl 2: II246-II255.

Clustering techniques for classification of splice sites of human exons

Muselli M (1), Romeo F (2), Pfeffer U (2)

(1) Institute of Electronics, Computer and Telecommunication Engineering, Italian National Research Council, Genova
(2) Functional Genomics, National Cancer Research Institute, Genova

Motivation

The usage of genetic information by the cellular machinery has been greatly facilitated by the evolution of splicing, a process that permits to join fragments of coding sequences (exons) from larger genomic regions containing prevalently non-coding sequences (introns). The process relies on the precision of exon recognition since a single nucleotide shift leads to an alteration of the reading frame and hence to altered information. The analysis of the splice sites has led to the identification of the consensus sequences GURAGU at the exon-intron border and AXnYnAG (branch point, polypyrimidine tract, AG) for the intron-exon border. However, these consensi are weak: many real sequences considerably divert from them, and a great number of sequences matching the consensus patterns are not used as splice sites. We follow the hypothesis that additional sequence features that contribute to the splice site definition can be identified if specific classes of such sites are considered. Yet there is no objective criterion to classify exons. We therefore set out to classify splice sites using machine learning approaches. The classification of splice sites can be used for analyses of correlation with the biological behavior of the relative exons such as alternative splicing. This approach is of foremost importance given the introduction of whole genome exon analyses by microarray hybridization. Class specific additional sequence features may yield new information on functional single nucleotide polymorphisms and somatic mutations that, without this information, would be considered as silent.

Methods

The classification of splice sites is performed by analyzing sequences of n bases around the transition points between exons and introns. In particular, the target of the analysis is to retrieve a specific characterization that distinguishes sequences including a splice site (denoted as positive sequences) from others that do not contain it (negative sequences). Any machine learning technique for classification can be adopted to deal with this problem; however, rule generation methods are to be preferred, since they are able to produce sequences in IUB code that detect the presence of the splicing site. Shadow Clustering (SC) [1,2] is a rule generation method, based on monotone Boolean function reconstruction, which is able to achieve performances comparable to those of best machine learning techniques. SC proceeds by grouping together binary strings that belong to the same class and are close to each other according to a proper definition of distance. Since SC operates on binary strings, every sequence of bases must be previously converted in Boolean form, before the generation of the set of rules starts. To this aim, the standard basis conversion: 'A' = '0111', 'C' = '1011', 'G' = '1101', 'T' = '1110' is employed. This gives rise to a training set for SC containing binary strings with length 4n. If a huge collection of negative sequences is included in the training set, the execution of SC generates a high number of rules, many of which are obtained through specializations of a general consensus pattern. To determine these relationships a proper hierarchical clustering technique is adopted; it can be viewed as a modification of the single linkage algorithm, which takes into account the presence of an ordering among the elements to be clustered, given by the relevance associated with each rule.

Results

DNA sequences extracted from the human genome have been considered for the classification of splice sites. Two different training sets have been taken into account: the first one concerns the detection of exon-intron (EI) transition points, whereas the second one is related to intron-exon (IE) sites. In both cases sequences containing n= 120 bases around the splice sites have been selected to be included in the training set as positive examples. In particular, by using the dataset of Clark and Tanaraj [3] a collection of 14,026 and 14,309 positive sequences have been generated for the EI and

the IE problem, respectively. The negative examples for the training set in both situations have been retrieved by analyzing a 300kB fragment of human genomic DNA extracted from the region on chromosome 6 containing the estrogen receptor ?, a gene that shows extensive alternative splicing [4,5] . In this way, the whole training set for the two problems contains more than 310,000 examples. The execution of SC has produced 2,618 and 3,064 rules, written as IUB code sequences, in the EI and in the IE problem, respectively. Most of them presents consensus patterns different from the standard GURAGU. The application of the hierarchical clustering approach, developed for this particular analysis, has recognized about 25 clusters for each problem, which can be associated with specific non standard consensus patterns that can be further examined to improve the understanding of the phenomena involved in the splicing mechanism.

Contact email: marco.muselli@ieiit.cnr.it

References

1. M. MUSELLI, A. QUARATI Reconstructing positive Boolean functions with Shadow Clustering. In Proceedings of the 17th European Conference on Circuit Theory and Design (ECCTD 2005), (Cork, Ireland, August 2005).
2. M. MUSELLI Switching neural networks: A new connectionist model for classification. In WIRN/NAIS 2005, vol. 3931 of Lecture Notes in Computer Science (2006) Eds. B. Apolloni, M. Marinaro, G. Nicosia, R. Tagliaferri, Berlin: Springer-Verlag, 23-30.
3. F. CLARK, T. A. THANARAJ Categorization and characterization of transcript-confirmed constitutively and alternatively spliced introns and exons from human. Hum Mol Genet, 11 (2002) 451-464.
4. U. Pfeffer, E. Fecarotta and G. Vidali. Coexpression of multiple estrogen receptor variant messenger RNAs in normal and neoplastic breast tissues and in MCF-7 cells. Cancer Res., 55, 2158-2165, 1995.
5. Ferro, P., Forlani, A., Muselli, M. and Pfeffer, U. Alternative splicing of the human estrogen receptor \alpha primary transcript: mechanisms of exon skipping. Int. J. Mol. Med., 12, 355-363, 2003.

Supplementary informations

Acknowledgment This work was supported by the "Ministero dell'Istruzione, dell'Università e della Ricerca" MIUR projects "Laboratory of Interdisciplinary Technologies in Bioinformatics (LITBIO)" and "Hormone Responsive Breast Cancer"

Getting the most out of comparative microarray data analysis: analysis of the estrogen-responsive transcriptome from breast cancer cells with four different microarray platforms

Mutarelli M (1,2), Scafoglio C (3,4), Cicatiello L (3,5), Colonna G (1),
Facchiano A (1,2), Weisz A (3,5)

(1) Centro di Ricerca Interdipartimentale di Scienze Computazionali e Biotecnologiche, Seconda Università degli Studi di Napoli
(2) Istituto di Scienze dell'Alimentazione CNR, Avellino
(3) Dipartimento di Patologia Generale, Seconda Università degli Studi di Napoli
(4) Department of Medicine, University of California S. Diego
(5) AIRC Naples Oncogenomics Center, Napoli (Italy)

Motivation

Presently, several commercial and academic providers offer printed DNA microarrays, also known as chips, prepared according to a variety of technologies. After a first generation of spotted cDNAs at a high density pattern onto a solid substrate such as a glass slide, the emerging standard is now the spotted or in-situ synthesis of short (25- to 30-mer) or longer oligonucleotides (50- to 70-mer) oligonucleotide probes directly onto a glass or silicon surface. A new and interesting innovation is the in-situ synthesis of the probes on beads, which are then randomly distributed on the chip surface. In order to evaluate the technical variability among different microarray platforms, we used four different commercial chips to study the gene expression profiles of hormone-responsive breast cancer cells following stimulation with estradiol. We decided to use only oligonucleotide-based arrays since previous comparative analysis have shown little reproducibility in cDNA microarray data and low overlapping in results among cDNA and oligo.

Methods

The following microarray platforms were used: i) the Affymetrix technology, based on 25 nucleotide-long oligonucleotides synthesized on a GeneChip® array, representing more than 39,000 transcripts derived from approximately 33,000 unique human genes; ii) the Agilent 'Human 1A Oligo' Microarray (G4110A) consisting of 60-mer, in situ synthesized oligonucleotide probes for a total of about 18000 different genes; iii) the Amersham CodeLink Human Whole Genome Bioarray, consisting of 30-mer, in situ synthesized oligonucleotide probes, for a total of about 52,000 different genes; iv) the Illumina Sentrix Human-6 (whole-genome) BeadArray, each containing 6 arrays which can be hybridized individually, consisting of 50-mer, synthesized in situ on beads randomly dispersed on the chip surface, for a total of about 46,000 different genes. The RNA derived from human breast cancer cells (ZR-75.1) stimulated for 72 hrs with 17β-estradiol (E2) after starvation in steroid-free medium for 4 days; the reference sample was derived from synchronized cells grown in steroid-free environment. A total intensity normalization was performed for the Agilent data, while the rma algorithm was used for the Affymetrix GeneChip expression data, and the same quantile normalization was performed for the Amersham Codelink data. As for the Illumina technology, since it was a new platform to our experience (and we had a lower number of technical replicates), we evaluated all four types of normalization provided by the Illumina BeadArray Image analysis software. Selection of significantly regulated genes was performed through the 'Significance Analysis of Microarrays' (SAM) software, setting the Delta value to gain a false discovery rate (FDR) of about 0.01 for all platforms. Particular care was used to find corresponding probes among platforms, since in our experience and in literature it has been shown that simple matching of probes using manufacturer's annotation (such as Unigene cluster or GeneBank accession number) could result in a little overlap of resulting gene sets and inconsistencies probably due to different design of probe sequences and/or non-updated annotation. For this reason, we decided to use probe sequences to verify the actual association to a well-curated sequence database as the NCBI RefSeq.

Results

Results will be presented in a poster and discussed will all participants during the meeting. Research supported by: Italian Association for Cancer Research (AIRC, Investigator Grants 2003), Italian Ministry for Education, University and Research: FIRB Post-genomica (Grant RBNE0157EH), European Commission (Contract BMH4-CT98-3433), Ministry of Health (Progetti Speciali 2000 and RF02/184), Second University of Naples (Ricerca di Ateneo 2002-2003).

Contact email: mailto:margherita.mutarelli@unina2.it

A computer model of X-inactivation

Nicodemi M (1), Prisco A (2)

(1) Dipartimento di Scienze Fisiche, Universita' di Napoli `Federico II', Via Cintia, 80126 Napoli, Italy
(2) Istituto di Genetica e Biofisica `A. Buzzati-Traverso', Via P. Castellino 111, 80131 Napoli, Italy

Motivation
Dosage compensation of X linked genes in female cells is a crucial process to survival and is achieved by the transcriptional silencing of one of their two X chromosomes, in many cases chosen at random. As many genetic and molecular aspects involved in X-inactivation are known, the very starting mechanism whereby cells count and chooses between two equivalent X chromosomes and randomly make a differentiating mark on only one of them is still not understood. The important scientific and medical implications of such a regulation mechanism have focused, in fact, substantial interest on its elusive origin.

Methods
We introduce a Statistical Mechanics inspired model of a “controlling factors” theory of X inactivation, which is investigated by computer simulations and checked against existing experimental evidence.

Results
Our model describes how the “blocking factor” complex is formed and how the symmetry in the binding of the complex to the equivalent X chromosomes is broken. In this way, it reconciles within a single framework the existing experimental evidences and points out that “counting” and “choice” are regulated by a single unifying mechanism. The simplicity and robustness of the regulation mechanism we illustrate for X-inactivation suggest it can underlay many cell processes involving allelic exclusion as well.

Availability: http://people.na.infn.it/~nicodem/

Contact email: Mario.Nicodemi@na.infn.it

FunGenAgent: An Agent-Based Approach for Workflow Composition in Homology Functional Genomics

Orro A (1,2), Armano G (3), Vargiu E (3), Milanesi L (1)

(1) CNR-ITB, Istituto di Tecnologie Biomediche-CNR, Segrate (Milano)
(2) CILEA, Consorzio Interuniversitario Lombardo per la Elaborazione Automatica, Segrate (Milano)
(3) Dipartimento di Ingegneria Elettrica ed Elettronica, Università di Cagliari, Cagliari

Motivation

Most tasks in bioinformatics analysis of genomics sequences cannot be carried out with a single standalone application. Most often, to solve a particular task, a combination of many computational tools and data sources is required. Due to the diversity in formats and interfaces and the low diffusion of standard methodologies for data exchange, the integration of heterogeneous computational and informative resources is a difficult task. In this work we present an agent-based approach aimed at supporting composition, execution and management of bioinformatics workflows. We also describe a preliminary implementation of this approach in the field of homology-based functional genomics.

Methods

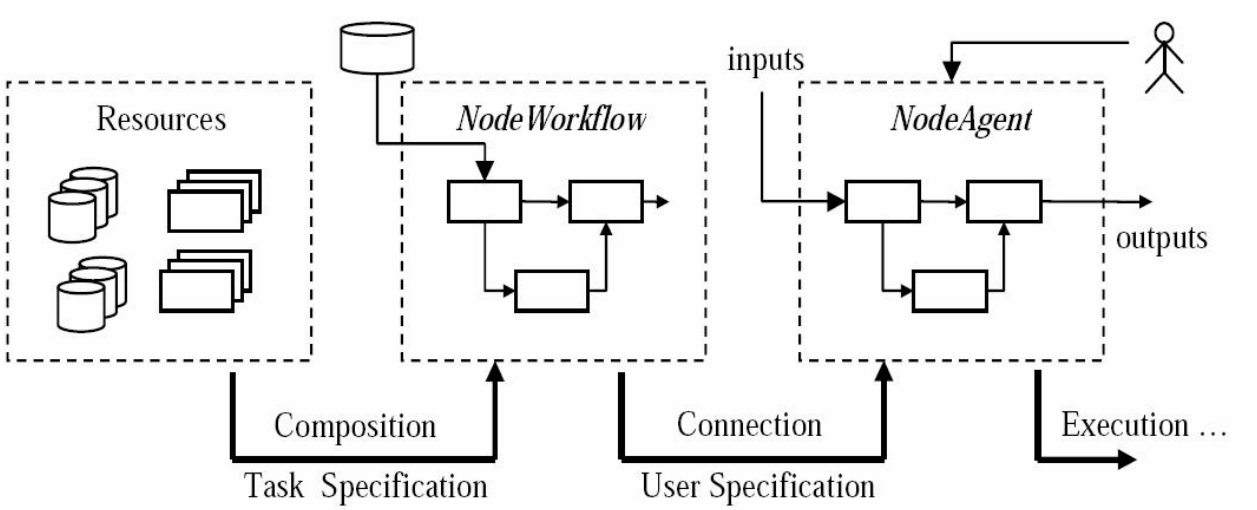
From an application-based point of view, workflows are considered as a network of nodes, each one aimed at performing a specific application. In this way, a user has to take into account all the details involved in each application. It is clear that, as the number of applications grows, this problem becomes more and more difficult to be tackled. To this end, we propose a task oriented view in which nodes are associated to a particular task, each task being assigned to a specific agent. In our scenario, a software agent (called TaskAgent) denotes a node of the workflow and exhibits a behavior that implements a specific application. In particular, each TaskAgent provides the available resources and a suitable user interface in order to satisfy the requirements of the task and the user, respectively (see the Figure). Furthermore, each agent exhibits additional features: (1) according to its domain knowledge, it selects suitable resource connections for the corresponding task, (2) it exports only those high-level parameters that are intuitive for the user; and (3) at execution time, it can interact with the user in order to monitor the overall process.

Results

A prototype of the proposed approach (called FunGenAgent) has been implemented and tested in functional genomics applications. The overall workflow is composed by three TaskAgents: (1) homology search handler, (2) multiple aligment handler, and (3) protein function predictor. The system gives as output a vector of class propensities for each class represented in the multiple alignment. Since the search of homologue sequences has been performed using the BLAST tool, here we focus only in the remaining steps. In both cases, the corresponding agent performs complex strategies in order to (i) integrate the information, (ii) execute, and (iii) combine the output of the involved application. The TaskAgent performs the multiple alignment embodies several multiple alignment tools. First, the multiple alignment is calculated with Clustal. Then, through a post-processing activity, the TaskAgent optimizes the alignment choosing the optimization strategies on the strength of the similarity of the sequences involved in the alignment. In fact, in low-similarity regions, a program for calculating multiple alignment by the secondary structure; whereas in the other regions a general multiple alignment optimization is more effective (in the current implementation, we adopt RASCAL). The TaskAgent that performs functional inference embodies several tools in order to assign a functional annotation to each part of the sequences. A set of distance between the target sequence and the protein is calculated. All distances weight the functional class taken from the function available from the database of functional family. The FunGenAgent prototype will be tested in the frame of the European Project BioinfoGRID http://www.itb.cnr.it/bioinfogrid (Bioinformatics Application for Life Science) and will be available

under the Italian MIUR-FIRB project LITBIO www.litbio.org (Laboratory for Interdisciplinary Technologies in Bioinformatics).

Contact email: alessandro.orro@itb.cnr.it



A new procedure to detect similarities among distant homologous proteins based on the comparison of domain flexibilities

Pandini A (1), Mauri G (2), Bonati L (1)

(1) Department of Environmental Sciences, University of Milano-Bicocca, Milano
(2) Department of Informatics, Systems and Communication, University of Milano-Bicocca, Milano

Motivation

Due to the improvement of molecular simulation codes and the availability of faster computational resources, a large amount of data on protein dynamics has been generated in the latest years. It is of great interest to investigate the informativeness of the data on dynamics and to effectively assess their usefulness in approaching and solving some classical problems of protein-protein comparison. When tertiary structures are available, the employment of structural alignment, supported by accurate statistical estimates, allows to detect similarities and derive accurate alignments. However the accuracy of the most sensitive methods remains comparable to reliable sequence-based methods, with a similar tendency in reporting false-positives in the case of difficult structural comparisons. This evidence suggests the need to employ some kind of additional information to improve protein comparison. Our proposal is that protein flexibilities derived from molecular dynamics simulations could be promising candidates in supporting detection of similarities among distant homologous proteins. This could be achieved through annotation of the dynamical properties of proteins with known structures and development of a fast and reliable procedure for large-scale comparisons. The proposed procedure could enhance homology search and alignment, as well as improve function detection and annotation.

Methods

The use of CONCOORD as a fast conformational sampling method is proposed and its reliability is assessed by comparing the results with those from Molecular Dynamics simulations. Essential Dynamics analysis is employed to extract a meaningful subspace of informative motions from the ensembles of structures generated by CONCOORD and the Root Mean Square Fluctuation (RMSF) of alpha carbons in the essential subspace is employed as a measure of the local flexibility of protein domains. On the basis of this representation, a synthetic index of similarity between domain flexibilities is proposed and the informativeness of this index is verified.

Results

To assess the reliability of the procedure, the dynamics of a collection of protein domains from a ASTRAL/SCOP40 fold is analyzed and the possibility to identify relationships at the family level on the basis of the dynamical features is discussed. The overall picture obtained is in good agreement with the SCOP classification, and suggests the presence of a distinguishable familiar trend in the flexibility profiles. Moreover the results obtained also support the complementarity of the dynamical and the structural information. The results of this first test support the hypothesis that the additional level of information provided by flexibility annotation, inaccessible by simple structural comparison, can be employed to detect functional similarities otherwise unrecoverable.

Contact email: alessandro.pandini@unimib.it

A fast conformational sampling approach to investigate enzymatic cold-adaptation mechanisms

Papaleo E (1), Pandini A (1), De Gioia L (2), Bonati L (1)

(1) Department of Environmental Sciences, University of Milano-Bicocca, 20126, Milan (Italy)
(2) Department of Biotechnology and Bioscience, University of Milano-Bicocca, 20126, Milan (Italy)

Motivation

The huge amount of sequence and structural data provided by genomic projects, opened new intriguing possibilities to understand at the molecular level the plasticity of life in adaptation to the so-called "extreme" environments, such as low temperatures at a molecular level. Comparisons among protein sequences or entire genomes from extremophilic and mesophilic organisms suggested possible adaptative strategies, confirming that there is no general solution to the achievement of structural and functional stability in extreme environments. In particular, the number of reports and the structural data on enzymes from psychrophilic organisms has increased significantly only recently, revealing that adaptative strategies are specific inside an enzymatic family and related to an increased molecular flexibility that in turn leads to increased catalytic efficiency and reduced stability. Computational methods which allow a good sampling of the protein conformational space, such as classical molecular dynamics simulation (MD), are a suitable tool to investigate structural mechanisms at the basis of enzymatic cold-adaptation, by comparing psychrophilic enzymes with the mesophilic homologues. However, in order to get reliable information, long simulation times are required, allowing the investigation of only few cold-adapted enzymes, generally in comparison with one mesophilic counterpart. The number of three-dimensional structures of cold-adapted enzymes is increasing as well as the number of mesophilic homologues that could be studied in order to carry out comparative analysis. Therefore, the employment of computational methods which are not time-consuming but which, at the same time, allow a conformational sampling comparable to MD, can be useful to clarify the details of the different molecular strategies used by different enzymatic families in order to face low temperature conditions.

Methods

In order to extend the previously performed analysis to a larger set of psychrophilic enzymes, CONCOORD was employed to sample the conformational space of whole families of cold-adapted proteins. CONCOORD is a fast and reliable method to generate an ensemble of conformations from a list of distance constraints derived from the starting structure. Each structure is generated by a random algorithm and optimised to satisfy the constraint list. A collection of 500-2000 structure is a good representative of the ensemble of accessible conformations in the neighbourhood of the starting structure. To this extent, the CONCOORD ensembles can be analysed with traditional Molecular Dynamics tools. Essential Dynamics analysis is particularly effective in defining and extracting a meaningful subspace of informative motions and it was employed in this study to highlight evolutionary adapted flexibilities in the structure of psychrophilic enzymes. Additionally a comparison of CONCOORD and MD results was exploited for some representatives with the aim of further validate the fast sampling in the specific case of psychrophilic proteins. MD simulations were carried out with GROMACS and encompassed a timescale of tens of nanoseconds. Independent runs with different initial distributions of velocities were performed in the NPT ensemble and explicit solvent.

Results

Preliminary results on some enzymatic families support the hypothesis that psychrophilic enzymes carried out different evolutionary strategies in order to cope with the detrimental effects of low temperature environments, among which structural flexibility emerges as the main adaptative character. Moreover, the employment of a fast sampling method appears to be suitable to perform

large screening of whole families before the application of more time-consuming simulation methods.

Contact email: elena.papaleo@unimib.it

Comparative molecular dynamics simulations of homologous enzymes to investigate enzymatic cold-adaptation: a family-centred point of view

Papaleo E (1), Riccardi L (1), Pasi M (1), Gonella Diaza R (1), Smalas AO (2), Brandsdal BO (2), Fantucci P (1), De Gioia L (1)

(1) Department of Biotechnology and Bioscience, University of Milano-Bicocca, Milan (Italy)
(2) The Norwegian Structural Biology Centre, Faculty of Science, University of Tromso, Tromso (Norway)

Motivation

In recent years, there has been increasing interest in the origin of enzymatic adaptation to low temperatures to understand both the protein folding and structure-function relationships, and for biotechnological and industrial applications. The number of reports on enzymes from cold adapted organisms has increased significantly over the past years, revealing that adaptative strategies varies among enzymes, which use different small selections of structural features in order to gain increased molecular flexibility that in turn leads to increased catalytic efficiency and reduced stability. Molecular flexibility and the characteristics related to cold-adaptation are often difficult to estimate using experimental methods, whereas molecular dynamics (MD) provides a suitable tool to evaluate flexibility and molecular properties of proteins and correlate them to their structural and functional features. In light of the above scenario, the systematic investigation of different enzyme families becomes crucial to unravel the cold-adaptation strategies discovered by specific families. In the present contribution we report an approach based on several long MD simulations of representative structures for mesophilic and psychrophilic homologous at different temperatures, to explore the molecular basis inside different enzymatic classes. The MD trajectories were compared and analyzed considering the time-evolution of different properties: secondary structure content, molecular flexibility indexes, intramolecular and protein-solvent interactions, solvent accessibility, molecular surroundings of selected residues, to electrostatic interactions and properties of the protein surfaces. This analysis lead to unravel putative structural and molecular determinants of thermolability, flexibility and activity at low temperatures for psychrophilic enzymes.

Methods

Several MD simulations were performed in the NPT ensemble, using the GROMACS simulation software package and allowing the collection of 36-48 ns trajectories for each system. The solvent was explicitly treated by periodic boundary conditions and the ionization state of charged residues was set to mimic a neutral pH environment. In order to sample efficiently the conformational space, independent MD simulations were carried out starting from the same atomic coordinates and using different initial velocities from a Maxwellian distribution. Multiple trajectories help to identify recurring features and to avoid artifacts arising from the simulation procedure. The trajectories were checked to assess the quality of the simulation using GROMACS routines. In particular, the root mean square deviation (rmsd) calculated with respect to the initial structure and detection of the structural clusters sampled during MD simulation by the Linkage algorithm allowed the determination of the stable portions of the trajectories, which were further analyzed using GROMACS tools, and the NACCESS, DSSP, DELPHI, PYMOL, VMD programs as well as suitable tools developed in our laboratory.

Results

The comparative MD approach reveals that modulation of the number of protein-solvent interactions is not the evolutionary strategy followed by the analyzed enzymatic families to enhance catalytic activity al low temperature. In addition, flexibility and solvent accessibility of the residues forming the catalytic sites are generally comparable in the cold- and warm- adapted enzymes. In some test cases, it turns out a localized flexibility and peculiar electrostatic surface properties or salt-bridge pattern, related to a particular amino acid composition, which clustered around the

functional sites of the psychrophilic enzymes. In contrast, the mesophilic counterparts show a scattered flexibility in non-functional regions and additional stability in the surroundings of the active site as well as specificity pocket due to the presence of unique stabilizing ion-pairs. The results support the hypothesis that flexibility is the main adaptative character of psychrophilic enzymes, being responsible for the decrease of activation enthalpy that leads to increased kcat values at low temperature. On the contrary, other cold-adapted enzymes have evolved weakening intramolecular interactions, increasing structural disorder, showing a less stable binding of ion cofactors, and therefore enhancing the structural flexibility of the main protein domains, indicating a strategy of increased overall flexibility. In conclusion, the present investigation contributes, by means of a suitable computational and comparative approach to the clarification and enforcement of the picture that, among the several putative mechanisms of molecular and structural cold-adaptation, different enzymatic families can pursue different strategies according to their function, cellular localization and specific substrate.

Contact email: elena.papaleo@unimib.it

IMAGE: a new tool for the discovery of Transcription Factors binding sites

Paparcone R (1), Casilli R (1), Melchionna S (2), Marongiu A (1,3),
Palazzari P (1,3), Rosato V (1,3)

(1) Ylichron Srl, c/o ENEA Casaccia Research Center, Via Anguillarese 301, 00060 S.Maria di Galeria (Roma)
(2) INFM-SOFT, Department of Physics, University of Roma "La Sapienza", P.le A.Moro 5, 00186 Roma
(3) ENEA, Portici Research Center, Computing and Networks Service, Via Vecchio Macello, 80055 Portici
(4)ENEA, Casaccia Research Center, Computing and Modelling Unit, Via Anguillarese 301, 00060 S.Maria di Galeria

Motivation
The discovery of Transcription Factor binding sites is still an open problem, as most of the softwares available to date have low predictive character, particularly for complex DNA (such as the human DNA). A novel method is proposed which overcomes some of the limitations affecting the existing prediction tools.

Methods
IMAGE strategy for the discovery of TF binding sites is based on a novel approach inspired by a technique used for lossy image compression, known as vector quantization and by analogous methods to identify genes with similar functions and reconstruct phylogenetic trees by clustering algorithms. The central idea is to map all possible n-length substrings of a given DNA sequence into a properly defined n-dimensional space equipped with a distance measure which projects similar substrings, representing the same motif, into nearby points. Consequently, the goal of finding recurrent similar strings is shifted into the determination of highly clustered data points.

Results
A complete assessment of the IMAGE tool has been provided by using the web service available at the Washington site (http://bio.cs.washington.edu/assessment/submit.html), where most of the available tools for the TF binding sites discovery have been recently compared. Results demonstrate the ability of IMAGE in correctly predicting a large number of motifs with respect to the other tools, although with a minor sensitivity in discriminating false positives.

Availability: http://image.ylichron.it/

Contact email: rosato@casaccia.enea.it

Motif based classification of coexpressed genes

Pavesi G (1), Valentini G (2), Mauri G (3), Pesole G (4)

(1) Dept. of Biomolecular Science and Biotechnology, University of Milan
(2) Dept. of Computer Science, University of Milan
(3) Dept. of Computer Science, Systems and Communication, University of Milano-Bicocca
(4) Dept. of Biochemistry and Molecular Biology, University of Bari

Motivation
Understanding the complex mechanisms regulating gene expression is one of the greatest challenges for molecular biology. In particular, transcription is modulated by the interactions of transcription factors (TFs) with short DNA regions (TFBS, i.e. transcription factor binding sites) that are recognized in a sequence specific manner. The availability of genomic sequences, together with data concerning the expression of genes, has opened new opportunities in this field. In this work, we focused on the two following problems related to gene expression regulation: a) assessing whether classes of functionally related genes may be predicted using information extracted from their promoter sequences; b) the selection of motifs (TFBS) mostly related to the gene functional classes.

Methods
A quite common approach to the identification of TFBS involved in the regulation of transcription is the extraction from the promoters of a set of co-regulated (or co-expressed) genes of one or more conserved motifs, likely to represent instances of conserved TFBSs recognized by the same TF(s). Other promoters, presenting instances of the same motifs can then be predicted to be regulated in a similar fashion. The problem is that often, in the absence of experimental validation, it is very hard to assess 1) whether the conserved motifs found actually correspond to functional sites 2) which, among many candidates, are truly responsible of the regulation of the genes and, 3) if the motifs deemed to be significant (or, even those that are experimentally validated) are sufficient to explain the co-expression of the genes.

Results
We present a classification algorithm that predicts whether a gene may belong to a given set of co-expressed genes using information extracted from its promoter sequence. The classifier needs a training set composed of a set of co-expressed genes (the positive set), and a negative set, comprising genes not related to the first group. In other words, the classifier determines, from the motifs detected in the promoter of the gene, whether it is likely to be co-expressed with the genes of the positive set or not. The algorithm is composed of two steps: motif extraction and training of a classifier. In the first, motifs are extracted from each of the promoters of the genes investigated. For this task, we employed the Weeder algorithm. The main advantage is that, being Weeder an exhaustive algorithm, the result of the motif extraction phase covers every possible motif of a chosen length, and avoids the need to introduce significance criteria to select which motifs have to be used for the classification of the sequences. Moreover, Weeder is able to process each sequence separately, producing, given a motif length m, a vector composed of 4^m motif scores for each of the promoter sequences of the training set. Then, in the second step, a linear classifier is trained on the obtained score vectors. The first experiments we performed have shown very promising results. On different clusters of yeast genes with testing performed with leave-one-out cross-validation the average prediction accuracy ranged between 75% and 85%, varying according to the different clusters examined. Also, simple feature selection methods applied to the trained classifier permitted the extraction of the most significant motifs, that is, motifs that gave the greatest contribution to the correct classification of the examples. In most of the cases, the motifs matched experimentally known yeast transcription factor binding sites responsible for the regulation of the genes. We are now extending the basic idea to multi-class prediction and to other species, also evaluating the improvements deriving from the use of more sophisticated machine learning methods. While it is

well known that in metazoan organisms promoter sequence alone is often not sufficient to fully explain the patterns of expression of a gene, knowing if, and especially in which cases it can be predicted with our technique could provide valuable information and insights.

Contact email: pavesi@dico.unimi.it

Mapping OMIM mutated residues on PDB protein structures

Peluso D, Via A, Ausiello G, Helmer-Citterich M

Centro di Bioinformatica Molecolare, Department of Biology, University of Tor Vergata, Rome

Motivation

The OMIM database (1) is a collection of hereditary point mutations associated to diseases in Homo sapiens. Such mutations have been recently mapped onto the Swiss-Prot sequences (2). Despite the increasing number of protein structures stored in the PDB database, a correspondence between OMIM mutated residues and PDB residues has not yet been established. This type of information could help in providing insight into the molecular mechanisms that cause hereditary diseases. In this work, we performed an accurate mapping of OMIM mutations onto 3D protein structures. The results of the entire procedure will be used to further annotate amino acids in the pdbFun (3) database and will be available through the resource web site (http://pdbfun.uniroma2.it/).

Methods

The Swiss-Prot sequence numbers of OMIM missense mutations, provided by Martin and co-workers (2), have been transferred to PDB structures via the seq2struct resource (4), which establishes reliable links between Uniprot sequences and PDB or SCOP structures. Swiss-Prot sequences have been aligned to the sequences extracted from the ATOM coordinates of the PDB files, by retaining only sequence-structure pairs displaying a sequence identity greater than 90%. A supplementary BLAST local alignment, based on more stringent thresholds, was performed in a short region including the mutated amino acid on the Swiss-Prot sequence, in order to obtain a punctual residue mapping.

Results

In a non-redundant PDB set of structures, we found 3376 mutations associated to 1308 mutation sites, which are linked to 178 OMIM entries. As soon as the mapping is complete, we intend to use the OMIM mapping on protein structures to explore and further analyze cases where the effects of a mutation onto a protein structure (might) account for the protein misfunction and, hence, for the associated disease (5). In the future, we want to apply the mapping procedure and the subsequent structural analysis also to SNPs (6).

Contact email: daniele@cbm.bio.uniroma2.it

References

1. Hamosh,A., Scott,AF., Amberger,J., Valle,D., McKusick,VA. (2000). Mendelian Inheritance in Man (OMIM). Hum Mutat.;15(1):57-61.
2. Andrew C. R. Martin. (2005). Mapping OMIM mutations to SwissProt . Bioinformatics, Application Note, Vol. 00 no.00 pages 1-2
3. Ausiello,G., Zanzoni,A., Peluso,D., Via,A., Helmer-Citterich,M. (2005). pdbFun: mass selection and fast comparison of annotated PDB residues. Nucleic Acids Res.;33(Web Server issue):W133-7.
4. Via,A., Zanzoni,A., Helmer-Citterich,M. (2005). Seq2Struct: a resource for establishing sequence-structure links. Bioinformatics;21(4):551-3. Epub 2004 Sep 28.
5. Yue,P., Li,Z. and Moult,J. (2005). Loss of Protein Structure Stability as a Major Causative Factor in Monogenic Disease. J. Mol. Biol. 353,495-473.
6. Wang,Z. and Moult, J. (2001). SNPs, Protein Structure, and Disease. Human Mutation 17:263-270.

Assembling the Yeast Interactome

Persico M, Kiemer L, Cesareni G

Department of Biology, University of Tor Vergata, Roma

Motivation
How is the yeast proteome wired? This important question is still unanswered in spite of the abundance of protein interaction data obtained using high-throughput approaches and made available to the scientific community. Unfortunately, such large-scale studies show remarkable discrepancies in their results and coverage so combining them is not a trivial task: information on interactomes, or networks of interacting proteins, produced with diverse experimental techniques,each with their own inherent experimental errors, must be evaluated to construct a trustworthy protein interaction network.

Methods
Although the task of integrating different data sources has already been undertaken by different groups, the recent availability of the results of two new large scale studies has motivated a fresh approach to the problem. We examine different algorithms and approaches to the problem and introduce our own model for building a trustworthy network of protein interactions.

Results
We present a draft of the yeast interactome taking advantage of various heterogeneous sources of data: tandem-affinity-purification coupled to mass spectrometry (TAP-MS) data, large-scale yeast two-hybrid studies, and literature data stored in dedicated databases of protein-protein interactions. We compare our interactome with others available and we evaluate it for biological consistency. A trustworthy interactome is the starting point of other kinds of analyses: it can be used in simulation studies (cell automata) aimed at discovering the dynamic and evolutionary properties of molecular networks, and it can be useful for pathway analysis and network reconstruction studies. Ultimately, it can be used in machine learning approaches to predict or evaluate new protein interactions.

Contact email: maria@cbm.bio.uniroma2.it

The genomic signature for in vitro-induced invasive growth is enriched in genes correlated with human cancer aggressiveness

Piccolis M, Medico E

The Oncogenomics Center, Institute for Cancer Research and Treatment, University of Torino

Motivation
Gene expression profiling has been extensively used to study human cancer and define gene signatures whose expression correlates with specific features of the tumor. However, such signatures generally lack biological insight, as gene selection is only based on correlation with clinical features of interest.

Methods
To build signatures with greater biological significance, we set-up a statistical procedure for meta-analysis of DNA microarray data named Signature Enrichment Analysis (SEA). SEA is aimed at assessing whether a gene expression signature defined in a given in vitro biological model (e.g. genes regulated by a growth factor) is significantly enriched in genes whose expression in tumours correlates with a specific clinical and/or pathological feature. If the enrichment is significant, such genes can be used to build a cancer signature encompassing both clinical relevance and biological meaning.

Results
We applied SEA to a signature composed of genes transcriptionally regulated in vitro during the induction of invasive growth by tyrosine kinase receptors for Hepatocyte Growth Factor (HGF) and Epidermal Growth Factor (EGF). This signature was tested against gene expression datasets of lung and prostate cancer. We found that the invasive growth signature is enriched in genes discriminating specific features of the cancers explored, and in particular: (i) lung cancer propensity to metastasize to brain or liver; (ii) Gleason score of prostate cancer; (iii) prostate cancer as opposed to normal prostate tissue. From this signature we derived and preliminarily validated a series of classifiers correlated with prostate and lung cancer aggressiveness. These results indicate that the genomic signature for in vitro-induced invasive growth captures a transcriptional program that can be reconstructed, albeit partially, in human cancer samples.

Contact email: enzo.medico@ircc.it

Genomic analysis of gene structure and expression of Plasmodium falciparum rifins

Pizzi E, Bultrini E, Silvestrini F, Alano P

Dipartimento di Malattie Infettive, Parassitarie e Immunomediate, Istituto Superiore di Sanità, Roma

Motivation

Rifins are members of the most abundant multigene family of Plasmodium falciparum genome. They are two-exon genes coding for transmembrane proteins that are probably located at the surface of infected erythrocytes where they have been supposed to contribute to antigenic variability of the parasite. Because of their subtelomeric location they are subjected to frequent recombination events leading to new repertoires of genes coding for proteins with novel antigenic properties. Although the completion of genome sequencing an extensive analysis on rifin organization and distribution has not been carried out to date, despite their probable important role in host-parasite interactions. In this work we present a sequence analysis of the entire repertoire of rifin genes in P. falciparum genome. We investigated 5' upstream and 3' downstream sequences for the presence of potential regulatory elements and for conformational and compositional properties. Then we analysed available gene expression data.

Methods

We extracted 5' upstream (500 bp), coding and 3' downstream (500 bp) sequences for each of the 150 rifin genes in the P. falciparum genome. Within each group of sequences a comparative analysis was carried out: each sequence was compared with all the others and a distance (D=100-percentage of identity) was calculated for each comparison. These data were used as input for a multidimensional scaling. This procedure corresponds to map sequences on a two-dimensional space in which relative distances are maintained and hence allowed us to easily recognize sub-families of them. According to these results we proposed a classification for rifin genes and on the basis of a simple probabilistic model we studied their distribution in the genome. Mean profiles for bendability propensity and nucleotide composition were constructed for 5' upstream and 3' downstream sequences. Further, we looked for common oligomers 8 bp long that are over-represented with respect a given background (lexicon-partitioning). We considered as significant only oligomers that occur at least in the 80% of sequences with an obs/exp ratio higher than 1.5.We analysed available gene expression data to identified rifin transcripts regulated during the asexual life cycle of the parasite.

Results

We applied several methods to characterize sequences corresponding to 5' upstream, coding and 3' downstream regions, to identify putative regulatory elements and to analyse available gene expression data. We found that 5' upstream sequences, as well as coding ones, can be grouped into two sub-families, whereas 3' downstream sequences are organized into three main clusters. On the basis of these results, we classified rifin genes as combinations of different 5', coding and 3' sequences. We found that only some arrangements of the three modules occur in the genome, some are completely absent and some others occur at the same frequency as the random expectation, suggesting positive selection acting on these genes. We investigated 5' upstream and 3' downstream sequences to look for potential regulatory elements and we found that in the case of 5' upstream sequences potential transcription start sites are probably located about at -200 bp from the first codon. Finally we studied available gene expression data. We found that only 7 out of 150 possible rifin transcripts are regulated during the asexual life cycle of the parasite, supporting the hypothesis that these genes are subjected to antigenic variation mechanisms.

Contact email: epizzi@iss.it

Expression levels and gene function influence transposon occurrence in mammalian introns

Pozzoli U (1), Menozzi G (1), Cereda M (1), Comi GP (2), Sironi M (1)

(1) Scientific Institute IRCCS E.Medea - Bioinformatics Lab.
(2) Dino Ferrari Centre, Department of Neurological Sciences, University of Milan, IRCCS Ospedale Maggiore Policlinico, Mangiagalli and Regina Elena Foundation , 20100 Milan, Italy.

Motivation

Transposable elements (TEs) represent more than 45% and 37% of the human and mouse genomes, respectively. Once considered as merely junk DNA, it is now widely recognized that interspersed repeats have been playing a major role in genome structure evolution. Several studies have suggested that TE integrations have been subjected to purifying selection to limit the genetic load imposed to their host. Yet, a comprehensive analysis of the forces driving TE insertion, fixation and maintenance within mammalian genes is still missing.

Methods

NCBI Reference Sequence genes were selected for human and mouse. Gene and intergenic sequences as well as intron/exon boundaries were derived from the UCSC genome annotation database (http://genome.ucsc.edu/). Gene counts were: 7695 and 5550, for human and mouse, respectively, accountiong for 81989 and 55553 introns. Mulispecies conserved sequences (MCS) were obtained using phastCons predictions, which are available through the UCSC database. Transposable elements were identified and categorized using the UCSC annotation tables that rely on RepeatMasker. Microarray data on expression levels in human and mouse tissues were derived from previous studies based on high-density oligonucleotide arrays (GNF Gene Expression Atlas 2). For human and mouse, SAGE libraries, were obtained from the SAGE Genie website (http://cgap.nci.nih.gov/SAGE). For each transcript entry in our databases we extracted a SAGE tag; tags were then matched to all RefSeq mRNAs and purged if they corresponded to more than one transcript. We than matched our tags to those in libraries, added all counts for libraries representing the same tissue type and converted absolute counts to relative tag counts (c.p.m.). Statistical analysis were performed using R. For lowess smooths, five robustifying iterations were always performed and a smoothing span of 0.5 was used. To allow empirical p value calculations, we performed 100 independent random data permutations.

Results

We had previously suggested that TE insertions might be constrained in human introns by the presence of conserved elements. We analyzed the distribution of different TE families, namely Alu, MIR, L1, L2, LTR and DNA transopsons in human/mouse introns. Correlation analysis revealed that, in both mammals, the frequency of all TE families negatively correlates with MCS density when introns containing at least one MCS are analyzed. We next wished to verify whether different TE families might be differentially represented depending on gene function. TE frequency varies with intron length, GC%, and, as shown above, MCS density. For each TE family we performed multiple regression analysis using intron GC%, intron length and conserved sequence length as covariates; the regression fit was used to predict the expected TE number per intron (nTEiexp). For each gene the TE normalized abundance (TEna) was calculated as follows: TEna=[sum(nTEi exp)-sum(nTEi obs)]/[sum(nTEi exp)+sum(nTEi obs)] where nTEiobs is the observed TE number per intron. Genes displaying TEna>0.5 or TEna<-0.5 were classified as TE-rich or -poor, respectively and significant gene ontology associations were retrieved. Genes involved in morphogenesis/development were found to be overrepresented in all TE-poor groups (as previously noticed for the HOX gene cluster); the same holds true for genes encoding transcription factors and cellular proteins involved in basic functions. Interestingly, we identified another group of genes that in both human and mouse, are over-represented among TE-poor genes: it is the case of hormones and ckemokines/cyotkines. This finding suggests that the majority of TE, including Alus and old

TEs (usually considered relatively benign dwellers of mammlian genomes), might exert disturbing effects on genes that require subtle tuning of expression levels. TEs have been reported to differentially associate with gene regions depending on expression levels. In order to address this issue we analyzed variation in Tena as a function of mean gene expression (obtained from both microarray and SAGE data). Lowess smooths were calculated for each TE family and compared to curves obtained from random data permutations. In both mammals a marked decrease in Tena is observed for genes above the 70th- 80th gene expression percentile. We next calculated the intronic to intergenic normalized frequency difference: again, for highly expressed genes and for all TE families, a decreasing trend is observed when frequency differences are plotted against gene expression. These data suggest that independently of gene isochore and repeat type, a TE exclusion from intronic regions is observed that depends on gene expression level. In summary our data indicate that, although different TE types might exert distinct effects on gene regulation, gene features such as intronic MCS density, gene function and expression have been playing a major role in governing TE fixation and maintenance in mammalian intronic regions.

Contact email: uberto.pozzoli@bp.lnf.it

Pattern Discovery: a web based interface for exhaustive analyses on multiple biological sequences

Raimondo E (1,2), Chiusano ML (2)

(1) PhD fellow in Computational Biology, Interdepartmental Research Center for Computational and Biotechnological Sciences, Second University of Naples, Naples, Italy
(2) Department of Structural and Functional Biology, University 'Federico II', Naples, Italy

Motivation

The search for common motifs in biological data is fundamental to find structural correlation that could be informative of functional and/or evolutionary relationships. In its simplest form, the Generic Pattern Discovery Problem on biological sequences can be formulated as the problem of finding all the patterns that occur in at least K sequences in a given sample of n elements. However, the relevant discovery problem is NP-hard. As a consequence, existing algorithms are commonly based on two main approaches: either they settle for incomplete results in order to achieve reasonable performance (approximation algorithms), or their execution time is suitable only for medium-sized inputs. We discuss here on a web based methodology to support pattern discovery in biological sequences. The algorithm proposed is based on a variant of the deterministic approach by Rigoutsos and Floratos (1). The web based approach aims to provide useful tools for suitable mining on multiple sequences. The novel algorithm is designed to overcome computational limits of an exhaustive pattern discovery approach.

Methods

We designed a novel algorithm for pattern discovery based on the TEIRESIAS method described in (1). TEIRESIAS solves the problem of finding all the maximal patterns occurring in at least K sequences in a set of n elements. Considering "maximal patterns" reduces the output redundancy as well as the computational complexity typical of an exhaustive search, while the "density" of the patterns can be user-driven by the input parameters L and W, where L indicates the minimum number of defined characters in every sub-pattern which length is at most W. A preliminary "scanning" phase determines all of the elementary patterns, i.e. patterns which length is at most W, with exactly L defined characters. Then, a combinatorial search for more specific patterns, starting from the set of elementary ones, follows. This step of the algorithm is termed the "convolution" phase. In our strategy, the scanning phase is the same as in (1), while the convolution step is rather different. Indeed, the main drawback in TEIRESIAS algorithm is the need of memory space when stacking all the possible extensions of elementary patterns in a representative set of long sequences. We re-designed the algorithm by splitting the convolution phase into a number of different steps, which, in their turn, are based on several convolve-send-receive cycles. Patterns generated at each step are used, during the following step, to generate new patterns, and then removed, as they won't be used any more. This leads to a more efficient data management and memory usage. A friendly PHP-based interface has been built to exploit the algorithm for different purpose in biological sequence analyses and to allow the software usage through web access.

Results

Our efforts have been focused on two principal aspects: a novel algorithm strategy and a suitable PHP-based interface that supports specific analyses on both nucleic acids and protein data. Our strategy is based on the re-organization of an exhaustive algorithm (1) to provide a more efficient implementation in terms of space requirements. Moreover, our method results suitable for a parallel implementation, improving time computational costs. The web based interface to the algorithm was designed to exploit the algorithm for solving varied specific analyses on biological sequence data as well as to allow user-driven mining on the reported results.

Contact email: enrico.raimondo@unina2.it

References
1. Rigoutsos, I., and Floratos, A. (1998) Bioinformatics 14(1):55-67.

Supplementary informations
This work is supported by the Agronanotech Project (MIPAF, ITALY)

Splicy: a web-based tool for the prediction of possible alternative splicing events from Affymetrix probeset data

Rambaldi D (1), Felice B(1), Praz V (2), Bucher P (2), Guffanti A (1)

(1) The IFOM-IEO Campus, Via Adamello, 16 - 20139 Milano, Italy
(2) ISREC, Ch. des Boveresses 155, Epalinges (Switzerland)

Motivation

The Affymetrix technology is nowadays a well-established method for the detection of gene expression profiles in cancer research studies. As an example, a query with the keyword 'cancer' of the NetAffx Scientific Publications database (http://www.affymetrix.com/community/index.affx) results in the retrieval of 672 publications from 1996. However, changes in gene expression levels are not the only aspect of importance in the link between genes and cancer. The existence of gene isoforms specifically linked with cancer or apoptosis is increasingly found in literature (1,2).

Methods

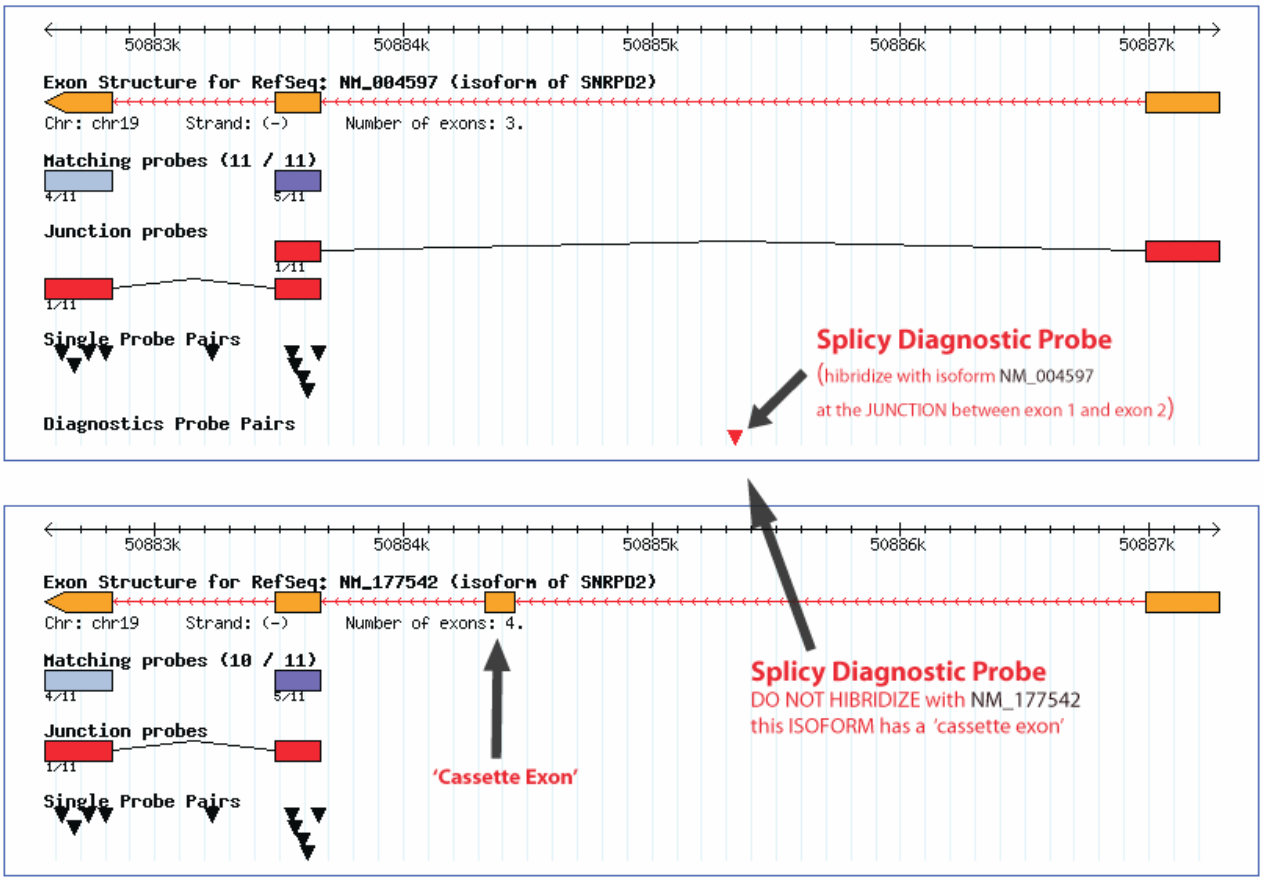
We present here a web-based software tool, Splicy (http://bio.ifom-ieo-campus.it/splicy/), whose primary task is to retrieve data on the mapping of Affymetrix probes to single exons of gene transcripts and display graphically this information on all available transcripts (both RefSeq and cDNAs). The program accepts in input a list of Affymetrix probesets and produces a series of graphical displays, each relative to a transcript associated with the gene targeted by a given probe. Each graphic reports the exon structure for a given transcript, a gliph which evidentiates the exons which contain matching probes, a gliph evidentiating the probes which are at the boundary of two exons, and a series of triangular glyphs evidentiating each single probe on the matching exon. If a given probe belongs to an exon which is skipped in a different transcript belonging to the same gene, it is tagged as a possible "splice diagnostic probe" and marked red. The idea is that a given probeset containing a 'diagnostic' probe will behave differently in the hybridization process, according to the transcript variant which is present in the hybridization mixture. The information on the transcript-by-transcript and exon-by-exon mapping can be retrieved both graphically and in the form of tab-separated files. Other features which can be retrieved from the graphic are the annotation table for each probeset, the Probeset design, RefSeq targets, the complete list of the Probeset probe pairs, the coordinates of the Genome alignments, further notes and links on additional transcript/gene annotation, GO functional classification and a direct link to the corresponding Entrez GENE entry. The mapping of single probes to RefSeq or EMBL cDNAs derives from the ISREC mapping tables which are the basis of the CleanEx Expression Reference Database Project (http://www.cleanex.isb-sib.ch/). We currently maintain mappings on the most popular human and mouse Affymetrix chips, and Splicy can be queried for matches with human and mouse RefSeq or EMBL cDNAs.

Results

We think that Splicy will be useful for giving to the researcher interested in transcriptome diversity a clearer idea of the possible transcript variants linked with a given gene and an additional key of interpretation of microarray experiment data. Splicy is publicly available from the web link http://bio.ifom-ieo-campus.it/splicy/ and has been realized thanks to part of a bioinformatics grant from the Italian Cancer Research Association.

Availability: http://bio.ifom-ieo-campus.it/splicy/

Contact email: davide.rambaldi@ifom-ieo-campus.it


50883k
50884k
50885k
50886k
50887k
Exon Structure for RefSeq: NM_004597 (isoform of SNRPD2)
Chr: chr19
Strand: (-)
Number of exons: 3.
Matching probes (11 / 11)
4/11
5/11
Junction probes
1/11
1/11
Single Probe Pairs
Diagnostics Probe Pairs
Splicy Diagnostic Probe
(hibridize with isoform NM_004597
at the JUNCTION between exon 1 and exon 2)
50883k
50884k
50885k
50886k
50887k
Exon Structure for RefSeq: NM_177542 (isoform of SNRPD2)
Chr: chr19
Strand: (-)
Number of exons: 4.
Matching probes (10 / 11)
4/11
5/11
Junction probes
1/11
Single Probe Pairs
'Cassette Exon'
Splicy Diagnostic Probe
DO NOT HIBRIDIZE with NM_177542
this ISOFORM has a 'cassette exon'

FPGA implementation of a greedy scheme for Bioinformatics applications

Rampone S (1), Aloisio A (2), Izzo V (2)

(1) Universita' del Sannio - Research Centre on Software Technology RCOST - Benevento, Italy
(2) Universita' di Napoli "Federico II" - Dipartimento di Scienze Fisiche and INFN - Napoli, Italy

Motivation

In the past decade there has been an explosive growth of biological data, including genome projects, proteomics, protein structure determination, cellular regulatory mechanisms, and the rapid expansion in digitization of patient biological data. Although raw computational power follows "Moore's Law", the genomic data at GenBank (the NIH genetic sequence database, an annotated collection of all publicly available DNA sequences) is doubling every six months. Proteomic and cellular imaging data appear to grow even faster. Post-genomic-era bioinformatics will require high-performance computing power of the order of several hundreds of teraflops or more. Moreover a growing number of related problems is posed as complex optimization.

Methods

In recent years, FPGAs, short for Field-Programmable Gate Arrays, logic programmable chips [1] [2] have emerged as high-performance computing accelerators capable of implementing fine-grained, massively parallelized versions of computationally intensive algorithms [3]. In particular several problems arising in biomedical and bioinformatics research can be viewed as finding the optimal covering of a finite set [4] [5]. While the Set Covering problem is known to be NP-complete [6] a number of approximation heuristics have been proposed. The most efficient schema remains the greedy one [7]. Recently, a new greedy algorithm for approximating minimum set cover has been presented [8]. The algorithm, while not randomized, is based on a probability distribution that leads the greedy choice. It shows very good empirical performances and it has successfully been applied in wireless network applications [9] [10]. While efficient implementations are given, the cost of probability distribution evaluation can still be unaffordable in massive real-time applications. In this paper we describe an implementation based on a FPGA of a tailored version of the algorithm. It makes the algorithm suitable for several real world bioinformatics problems.

Results

The test results show very good empirical performances on the used benchmarks. The speed up of our approach is also successfully tested.

Contact email: rampone@unisannio.it

References

1. B. L. Hutchings, and M. J. Wirthlin, "Implementation approaches for reconfigurable logic applications", inW. Moore andW. Luk, editors Field- Programmable Logic and Applications, Oxford, Springer Verlag, 1995, pp. 419-428.
2. W. Mangione-Smith, B. Hutchings, D. Andrews, A. DeHone, C. Ebeling, R. Hartenstein, O. Mencer, J. Morris, V. Prasanna, and H. Spaanenburg, "Seeking solutions in configurable computing", IEEE Computer, December, 1997, pp. 38-43.
3. Keith Regester, Jong-Ho Byun, Arindam Mukherjee, and Arun Ravindran, Implementing Bioinformatics Algorithms on Nallatech-Configurable Multi-FPGA Systems, Xcell Journal, First Quarter 2005
4. Alexander Genkin, Casimir A. Kulikowski, Ilya Muchnik Set covering submodular maximization: An optimal algorithm for data mining in bioinformatics and medical informatics Journal of Intelligent and Fuzzy Systems, Special Issue: Challenges for future intelligent systems in biomedicine, Volume 12, Number 1 / 2002, Pages: 5 - 17

5. Jie Zheng, Timothy J. Close, Tao Jiang and Stefano Lonardi, Efficient selection of unique and popular oligos for large EST databases Bioinformatics Vol. 20 no. 13 2004, pages 2101-2112
6. M.R. Garey, and D.S. Johnson, Computers and Intractability: A Guide tothe Theory of NP-Completeness, New York, NY: W.H.Freeman, 1979.
7. D.S. Johnson, "Approximation Algorithms for Combinatorial Problems", Journal of Computer and System Sciences, 1974, pp. 256-278.
8. S. Rampone, "Probability-driven Greedy Algorithms for Set Cover", in Proc. VIII SIGEF Congress "New Logics for the New Economy", Naples, Italy, September, 2001.
9. S. Dhar, M.Q. Rieck, S. Pai, and E.J. Kim, "Various Distributed Shortest Path Routing Strategies for Wireless Ad Hoc Networks", in Proc. 5th Int. Work. on Distributed Computing - Lecture Notes in Computer Science, 2918, Springer Verlag, 2003.
10. S. Dhar, M.Q. Rieck, S. Pai and E.J. Kim, "Distributed Routing Schemes for Ad Hoc Networks Using d-SPR Sets", Journal of Microprocessors and Microsystems, Special Issue on Resource Management in Wireless and Ad Hoc Mobile Networks, vol. 28(8), 2004, pp. 427-437.

Correlation analysis of gene expression time series at multiple scales: from the entire genome to metabolic and signalling pathways

Remondini D (1,2,3), Neretti N (1,4), Milanesi L (5), Tatar M (6), Sedivy JM (7), Franceschi C (1,8), Castellani GC (1,2,3,4)

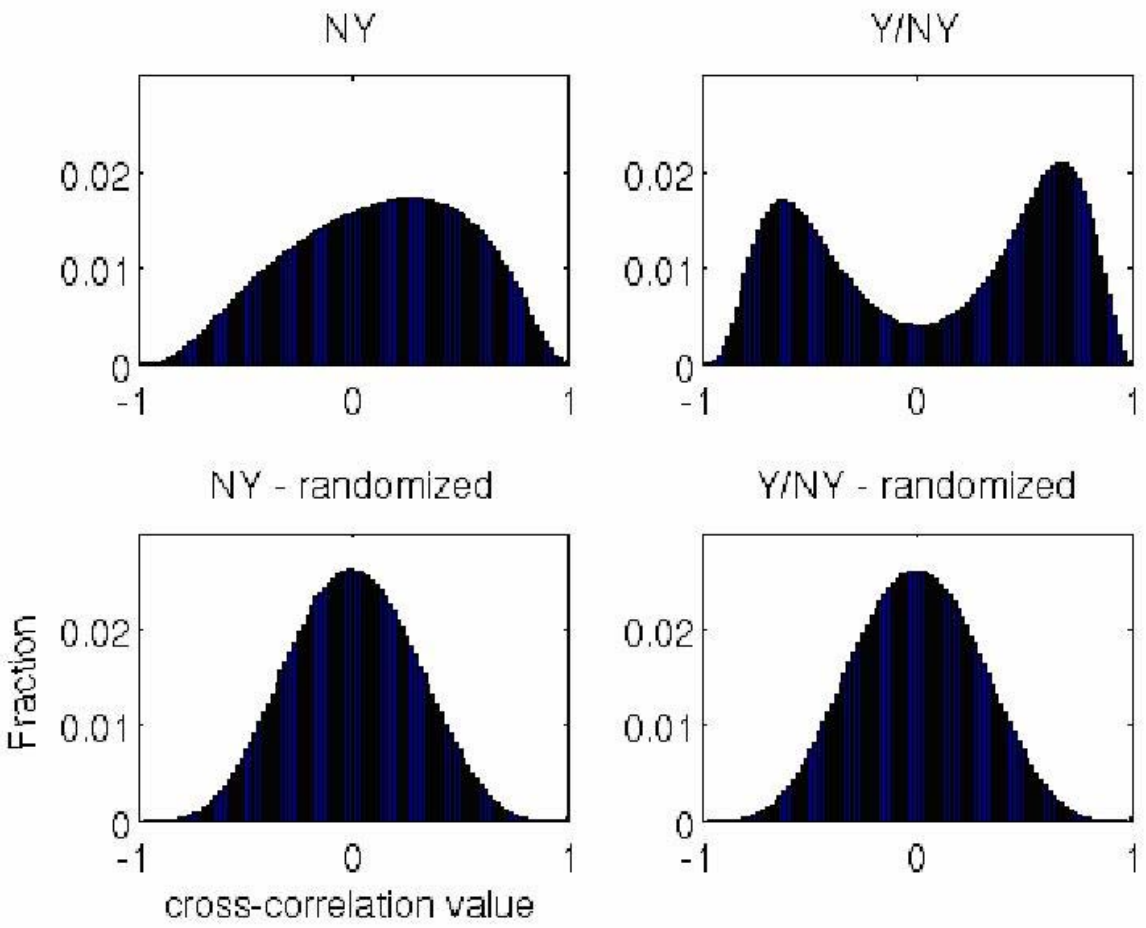
(1) Centro Interdipartimentale "L. Galvani", Università di Bologna IT
(2) DIMORFIPA, Università di Bologna IT
(3) INFN sezione di Bologna, IT
(4) Brown University, Providence RI USA
(5) Istituto di tecnologie biomediche (ITB) del CNR, Milano IT
(6) Dept. of Ecology and Evolutionary Biology, Brown University, Providence RI USA
(7) Dept. of Molecular and Cell Biology and Biochemistry, Brown University, Providence RI USA
(8) Dip. di Patologia Sperimentale, Università di Bologna IT

Motivation
High-throughput genomic data (microarray) can be very informative on cell state, but an emerging challenge is to retrieve useful informations about gene-gene interaction network from gene expression dynamics, obtained by array sampling collected over time. We propose a method based on the similarity between gene expression dynamics following a cell perturbation. Three examples are considerd: the characterization of the regulatory cascade of c-myc proto-oncogene following Tamoxifen stimulation in engineered rat fibroblas cells; the same characterization of genomic response in Drosophila after nutrition changes; patterns of gene activity as a consequence of ageing occurring over a life-span time series (25y-90y) sampled from T-cells of human donors.

Methods
The key for extracting useful gene network features both at a global level (quantifying the cell response to perturbation) and at a single gene level (gene targeting) is the correlation between expression time series. Thresholding methods are applied for noise removal and for network reconstruction, together with a (optional) processing that allows a preliminar selection of the genes possibly responding to perturbation.

Results
The method is applied to the different datasets: The correlation-based model can establish a clear relationship between network structure and the cascade of activated genes, reflected on multiple scales, from the whole genome down to pathways. The method results very sensitive to the temporal structure of the data, since data shuffling destroys the observed relations.

Contact email: gastone.castellani@unibo.it


NY
Y/NY
NY - randomized
Y/NY - randomized
Fraction
0.02
0.01
0
-1
0
1
cross-correlation value

Biowep: a workflow enactment portal for bioinformatics applications

Romano P (1), Bartocci E (2), Bertolini G (3), De Paoli F (3), Marra D (1),
Mauri G (3), Merelli E (2), Milanesi L (4,5)

(1) National Cancer Research Institute, Genoa
(2) University of Camerino, Camerino (MC)
(3) University of Milan Bicocca, Milan
(4) National Research Council, Milan
(5) CILEA, Segrate (MI)

Motivation

The huge amount of biological information, its distribution over the Internet and the heterogeneity of software tools that are used in bioinformatics makes the adoption of new data integration and analysis network tools a necessity. Information and Communication Technologies (ICT) standards and tools, like Web Services (WS) and Workflow Management Systems (WMS), can support the creation and deployment of such systems. WS are network services usually communicating by using the Simple Object Architecture Protocol (SOAP), a framework for the distribution of XML structured information, over HTTP. They offer a standardized programming interface so that software tools can effectively make access to the information and services they are delivering. Hence, they allow software applications to identify and interpret the information and, when ontological metadata is added, the associated semantics. WS have been implemented at bioinformatics centers; examples are Entrez Utilities at NCBI and the SoapLab implementation at the EBI, through which all EMBOSS programs are available. Workflows, defined as "computerized facilitations or automations of a business process, in whole or part" (Workflow Management Coalition, WfMC), aim to implement data analysis processes in standardized environments. Their main advantages relate to effectiveness, reproducibility, reusability of intermediate results and traceability. Some WMS have been proposed in bioinformatics. The Bioinformatic Workflow Builder Interface - BioWBI, Pipeline Pilot and Taverna Workbench [1] are the most known. These WMS assume that end users know which bioinformatics resources can be reached through a programatic interface and that they are skilled in programming and in building workflows, but they are not viable to the vast majority of researchers that are customised to web interfaces. A portal enabling the vast majority of unskilled researchers to take profit from these new technologies is still missing. We present here a web system that can support selection and execution of a set of predefined workflows. It presents a user-friendly web interface that is able to simplify access to such workflows and it therefore is viable to all end users.

Methods

The conceptual architecture of our system includes a Workflow Manager (WM), a User Interface (UI) and a Workflow Executor (WE). The WM is external to the prototype. Its task is the creation of predefined annotated workflows. These can be created by using different WMS. Presently, we allow for two of them: the Taverna Workbench [1] and the BioWMS [2]. Workflows enactment is carried out by the FreeFluo tool for Taverna workflows and by BioAgent/Hermes [3], a mobile agent-based middleware for the design and execution of activity-based applications in distributed environments, for BioWMS ones. The main processing steps of each workflow are annotated on the basis of their input and output data, elaboration type and application domain. Annotations are defined by using a classification of bioinformatics data and tasks. The UI supports end users authentication and profiling, including the classification of users on the basis of their job/role and scientific interests. Workflows can be selected on the basis of users' profiles. Available workflows can be searched through their annotation. Results of the execution of workflows can be saved and later analysed and possibly reused.

Results

We designed a web based client application, as defined in the WfMC Reference Model, that allows for the selection and execution of a set of predefined, annotated workflows. A prototype system is available on-line. It includes workflows that are devoted to the retrieval of data from IARC TP53 Mutation Database and from CABRI biological resources catalogues. Some of them have been made available both in Taverna and in BioWMS formats. Performances of the two approaches are under evaluation. The development and implementation of WS allowing the access to an exhaustive set of biomedical databases and analysis software and the creation of effective workflows through widely distributed WMS can significantly improve automation of in-silico analysis. biowep is available for interested researchers as a reference portal. They are invited to submit their workflows for insertion in the workflow repository. biowep is further being developed in the sphere of the Laboratory of Interdisciplinary Technologies in Bioinformatics - LITBIO.

Availability: http://www.o2i.it:8080/biowep/index.jsp

Contact email: paolo.romano@istge.it

References

1. T. Oinn, M. Addis, J. Ferris, D. Marvin, M. Senger, M. Greenwood, T. Carver, K. Glover, M. R. Pocock, A. Wipat and P. Li, Taverna: a tool for the composition and enactment of bioinformatics workflows. Bioinformatics, 20(17):3045-3054, 2004

2. Bartocci E., Corradini F., Merelli E., Scortichini L., BioWMS: a web based Workflow Management System for Bioinformatics. Submitted to BITS 2006.

3. F. Corradini and E. Merelli. Hermes: agent-base middleware for mobile computing. In Mobile Computing, volume 3465, pages 234-270. LNCS, 2005.

Supplementary informations

This work was partially supported by the Italian Ministry of Education, University and Research (MIUR), projects "Oncology over Internet (O2I)" and "Laboratory of Interdisciplinary Technologies in Bioinformatics (LITBIO)". Our system is partially based on open source. biowep is itself available under the GNU Lesser General Public Licence (LGPL). The portal is still under test.

Links Taverna Workbench: http://taverna.sourceforge.net/

BioWMS: http://litbio.unicam.it:8080/biowms/ (demo under testing)

BioAgent: http://www.bioagent.net/ Hermes: http://hermes.cs.unicam.it/

Precedence temporal networks for the representation of temporal relationships in gene expression data

Sacchi L, Larizza C, Magni P, Bellazzi R

Dipartimento di Informatica e Sistemistica, University of Pavia, Pavia

Motivation

The possibility of a genome-wide measure of gene expression offered by the DNA microarray technology is stimulating researchers in dealing with complex scientific challenges. Among them, the reconstruction of gene regulatory networks is particularly interesting. While in functional genomics genes are analyzed as single units to determine new functional information, in gene network reconstruction the interactions between genes are considered in order to infer regulatory connections between single genes or groups of them. In the field of discovering regulatory pathways, research is mainly focused on the analysis of gene expression time series collected by repeating DNA microarray experiments at different points in time. Many approaches based on different techniques of gene interaction representation have been proposed in the literature to deal with the topic of reconstructing gene regulatory networks. The observation that the dynamics of biochemical reactions may not be exactly captured by the (low) sampling time available in DNA microarray experiments has recently led to the introduction of the so-called module networks, where patterns of synchronized gene expressions are introduced. The present work introduces Precedence Temporal Networks (PTN), a novel method to extract from data and graphically represent temporal relationships between genes. Precedence Temporal Networks are a special kind of temporal network, where nodes (genes) are represented by properly defined temporal events while edges identify temporal relationships between the nodes.

Methods

The proposed approach develops through three steps: significant events are first identified in the time series by exploiting a qualitative representation of the profiles based on the technique of Temporal Abstractions (TAs). Precedence and synchronization relationships between abstractions are then searched through a set of properly extracted temporal rules; the resulting relationships are finally mapped into a labeled graph, the Precedence Temporal Network. In more detail, a first simple qualitative representation of the profiles made up of a set of consecutive basic trend TAs is obtained by processing raw data with a suitable algorithm. This description constitutes the starting point for the creation of the complex temporal events, i.e. complex abstractions which describe specific interesting temporal behaviors (typically user-defined) occurring in the data. A complex temporal event holds over an interval and is labeled through a set P={p1,…,pn}, where each pi represents an interesting qualitative behavior and is made up by the composition of simple labels of the kind Increasing, Decreasing, Steady (e.g. Increasing-Decreasing). Precedence and synchronization relationships are then selected through a method for temporal rules extraction that works looking for both the members of the rule coming from the set of the complex temporal events. In a PTN each gene complex behavior will be described by a node; edges will represent the relationships between the nodes extracted by the temporal rules. To specify both synchronization and precedence relationships we distinguish between two types of connections: the first, called co-occurrence connection, links elements characterized by the simultaneity of their temporal events, while the second, called precedence connection, corresponds to a precedence temporal relationship. The result will be a graphical representation of temporal relationships of the kind "an increasing-decreasing pattern in Gene A PRECEDES an increasing-steady pattern of Gene B".

Results

The method has been applied to the analysis of the expression of a subset of genes involved in human cell cycle regulation; in particular, we have analyzed 20 time series of 47 samples corresponding to well-studied genes known to show an expression peak in specific phases of the

cell cycle. The extracted network is able to clearly highlight the different cell cycle phases, and in particular the strong synchronization occurring at phase G1/S boundary and the weaker synchronization of the latter phases. The proposed approach is a novel contribution in the visualization of temporal relationships between a set of variables and its application in the analysis of gene expression data seems useful to summarize cellular behavior. Our approach is different from other methods, since it is devoted to describe precedence and synchronization of gene temporal patterns in a data set. This technique is still under development; future work will be mainly focused on the study of the formal properties of the network, and its application to more complex biological problems.

Contact email: lucia.sacchi@unipv.it

p53FamTaG : a database resource of human p53, p63 and p73 direct target genes combining in silico prediction and microarray data

Sbisà E (1), Catalano D (1), Gisel A (1), Grillo G (1), Licciulli F (1), Turi A (1), Liuni S (1), Pesole G (1,2), De Grassi A (2), Caratozzolo MF (2), D'Erchia AM (1,2), Navarro B (1), Tullo A (1), Saccone C (1,2)

(1) Istituto di Tecnologie Biomediche- Sede di Bari, CNR, , Via Amendola, 122/D 70126 Bari, Italy
(2) Dipartimento di Biochimica e Biologia Molecolare, "Ernesto Quagliariello", Università degli Studi di Bari, Via Orabona, 4, 70126 Bari, Italy

Motivation
The p53gene family is composed of three genes, p53, p63 and p73, with polyhedrical functions in pivotal cellular processes such as DNA synthesis and repair, growth arrest, apoptosis, genome stability, angiogenesis, development and differentiation. p53, p63 and p73 encode sequence-specific nuclear transcription factors which recognise the same responsive element (RE), but with a degree of specificity for the target genes that is quantitatively distinct. The three genes are differentially regulated and carry out specialized, non-overlapping functions. Their inactivation or aberrant expression may determine tumour progression or developmental disease. The discovery of several protein isoforms with antagonistic roles which are produced by the expression of different promoters and alternative splicing, widened the complexity of the scenario of the transcriptional network of the p53 family members. Therefore, the identification of the genes transactivated by p53 family members is crucial to understand the specific role for each gene in cell cycle regulation. To identify new direct target genes, we combined a genome-wide computational search of p53 family REs and microarray analysis. The huge amount of biological results produced raised a critical need for bioinformatic instruments able to manage and integrate the data and facilitate their retrieval and analysis. We have developed the p53FamTAG database (p53 FAMily TArget Genes), which contains p53 family direct target genes selected in the human genome searching for the presence of the REs and the expression profile of the target genes obtained by microarray experiments.

Methods
The genome-wide computational analysis was performed by using PatSearch, a pattern matching program implemented in DNAfan tool (DNA Feature Analyzer) developed in our Lab. These data were integrated with the microarray results produced in our Lab from the overexpression of different isoforms of p53, p63 and p73 stably transfected in isogenic cell lines, allowing to study in a comparable way the transcriptional activity of all the proteins in the same cellular background. p53FamTAG is a relational database, designed in a modular way so that, new data coming from different public resources and experimental analyses can be integrated and updated independently when needed. The p53FamTAG was implemented using MySQL as DBMS and the query/retrieval system was built using PHP Seagull Framework.

Results
p53FamTAG represents a unique integrated resource of human direct p53 family target genes, linked to other public databases (HUGO, EnsEmbl, RefSeq), and provides the user with an efficient query/retrieval system which allows the export of the RE sequences and results of our microarray experiments. p53FamTAG also contains 83 experimentally verified p53 family target genes, and 341 p53REs recently identified by ChIP-PET strategy linked to the relevant UCSC genome tracks and to PubMed entries. The database was developed for supporting and integrating high-throughput in-silico and experimental analyses and represents an important reference source of knowledge for research groups involved in the field of oncogenesis, apoptosis and cell cycle regulation.

Contact email: elisabetta.sbisa@ba.itb.cnr.it

Stem-loop structure search

Scalabrin S (1), Policriti A (1), Morgante M (2)

(1) Dipartimento di Matematica ed Informatica, Universita' di Udine, Udine
(2) Dipartimento di Scienze Agrarie ed Ambientali, Universita' di Udine, Udine

Motivation
The base-pairing of a nucleic acid secondary structure is a sort of biological palindrome. The base pairs of nucleic acid stems nest in a palindromic fashion with complementary base pairings rather than identical letters. In addition, nucleic acid stems are usually separated by a loop, i.e. a non palindromic sequence. Biological examples of such a structure can be found in hairpins at the 3' of helitrons, MITEs, microRNAs and tRNAs. At the moment, the problem of finding such a structure is solved through dynamic programming or in linear time with the use of the lowest common ancestor preprocessing.

Methods
We propose an algorithm which makes use of matching statistics on a suffix tree, is still linear on the size of the input string S, does not use the lowest common ancestor preprocessing and builds the suffix tree only for S and not for its complement reversed string, resulting in a great saving of memory, highly valuable in long sequence scan. Stems are easily detected using a suffix tree. Successively, the loop constraint (maximum distance between two stems) is solved linearly, on the number of putative stems, sorting the lists of occurrences in the input string S and its complement reversed string, and exploiting the linearity of the constraint.

Availability: Contact the authors via email to receive a copy of the software

Contact email: scalabrin@dimi.uniud.it

Introns containing conserved elements are evolutionary preserved in size

Sironi M (1), Menozzi G (1), Fumagalli M (1), Comi GP (2), Pozzoli U (1)

(1) Scientific Institute IRCCS E.Medea - Bioinformatics Lab. Bosisio Parini
(2) Dino Ferrari Centre, Department of Neurological Sciences, University of Milan, IRCCS Ospedale Maggiore Policlinico, Mangiagalli and Regina Elena Foundation , 20100 Milan, Italy.

Motivation

Different explanations have been proposed to account for within-genome intron size variations. Since we have previously demonstrated that fixation of multispecies conserved sequences (MCSs) influences intron size in humans, we wished to analyze whether this finding might be exploited to define a model for intron size evolution in mammals.

Methods

All genomic sequences were obtained from the UCSC genome annotation database (http://genome.ucsc.edu/). Only NCBI Reference Sequence genes were selected for human and mouse. Gene counts were: 7695 and 5550, for human and mouse (81989 and 55553 introns), respectively. Chimpanzee and rat genomic sequences were derived from UCSC; human/mouse mRNA alignments onto chimpanzee/rat genomic sequences were retrieved from the UCSC genome annotation database. Orthologous introns were then aligned using EMBOSS Stretcher with penalties of 48 and 1 for gap opening and extension, respectively. Human sequences mapping to gaps on the chimpanzee genome and covered by a Transposable element (TE, at least 85% of bases in TE and TE sequence extending no more than 10 bp 3' and 5' the gap) were considered human-specific insertions. Conversely, gaps longer than 80 bp in the human genome that were not accounted for by TE insertions or microsatellites in chimpanzee were classified as deletions. The same procedures were applied to mouse/rat intron alignments. For the identification of human-mouse and human-chicken orthologous pairs, the EnsMart database was interrogated and only entries representing unique best reciprocal hits (UBRH) were selected. MCS were obtained using phastCons predictions (available through the UCSC database). Transposable elements were identified and categorized using the UCSC annotation tables that rely on RepeatMasker. All statistical analysis were performed using R (http://www.r-project.org/). For lowess smooths, five robustifying iterations were always performed and a smoothing span of 0.5 was used.

Results

In order to study the impact of MCS presence on intron size, we aligned 27008 mouse-rat orthologous intron pairs and recorded mouse TE insertions and deletion events. Intron length before mouse-rat divergence (original length) was then estimated. In the case of mouse, TE insertion and deletion frequencies were analyzed after dividing introns in 4 length classes and, independently, in 4 groups depending on MCS density. Deletion frequency increase with intron size but, within each length class, both deletion and TE insertion frequencies diminish at the increase of MCS content (Kruskall-Wallis $p< 0.01$ for differences within all size groups). Interestingly, the deletion/insertion frequency ratio also decreases with MCS density increase (Kruskall-Wallis $p< 0.01$ for the second and third length classes). Analysis of the net variation in mouse intron size relative to the common mouse-rat ancestor indicated an average shrinkage, stronger for longer introns; yet, size contraction is progressively reduced, within each length class , for increasing MCS densities (Kruskall-Wallis $p< 0.01$), with introns being almost invariant when extremely rich in conserved sequence. For primate introns, we identified 280 TE insertions and 592 deletions in 39323 introns. Given the paucity of events and in order to analyze deletion and TE insertion frequency distribution as a function of MCS content, we applied a simulation based approach. For both insertions and deletions (separately) 1000 independent simulations were performed by randomizing TE insertion positions or deletion events. In the case of deletions we did not allow resulting introns shorter than 25 bp. For each intron, simulated TE insertion and deletion frequencies were assumed to conform to a Poisson

distribution and lambda was calculated. The Wilcoxon test for paired samples was used to compare the observed frequency with the expected (lambda). MCS-containing introns displayed significantly ($p<0.001$) less TE insertions compared to simulations. The same result applied to deletion frequency (paired Wilcoxon test, $p<0.001$). Conversely, MCS-lacking introns had significantly (paired Wilcoxon test, $p<0.001$) higher TE insertion and deletion frequencies. Consistent with the above findings is the analysis of normalized size variation (Δlength = [human - mouse length]/[human + mouse length]) in human-mouse orthologous intron pairs; lowess curves indicated that MCS-containing introns are extremely conserved in length irrespective of their size; conversely intronic regions lacking MCSs display remarkable size conservation until below length values ~1 kb and then diverge rapidly. Stronger size conservation for MCS-containig vs MCS-lacking introns also occurs over longer evolutionary periods as we were able to show by the analysis of human-chicken orthologous intron pairs. These results clearly indicate that MCSs have been posing a strong constraint to intron size evolution

Contact email: manuela.sironi@bp.lnf.it

The Genopolis Microarray Database

Splendiani A (1), Brandizi B (1), Even G (2), Ottavio B (2), Pavelka N (2), Pelizzola M (2), Mayhaus M (2), Foti M (2), Mauri G (1), Ricciardi-Castagnoli P (2)

(1) Dept. of Informatics, Systemistics and Communication, University of Milano-Bicocca.
(2) Genopolis Consortium, Dept. of Biotechnology and Bioscience, University of Milano-Bicocca.

Motivation
Gene expression databases are key resources for microarray data management and analysis. Public repositories as well as microarray database systems that can be implemented by single laboratories exists. However, there is not yet a tool that can easily support a collaborative environment where different users with different rights of access to data can interact to define a common content. The scope of the Genopolis database is to provide a resource that allows different groups performing microarray experiments related to a common subject to create a common coherent knowledge base and to analyze it, while respecting confidentiality of information. The Genopolis database has been implemented as a dedicated system for the scientific community studying dendritic and macrophage cells functions and host-parasite interactions

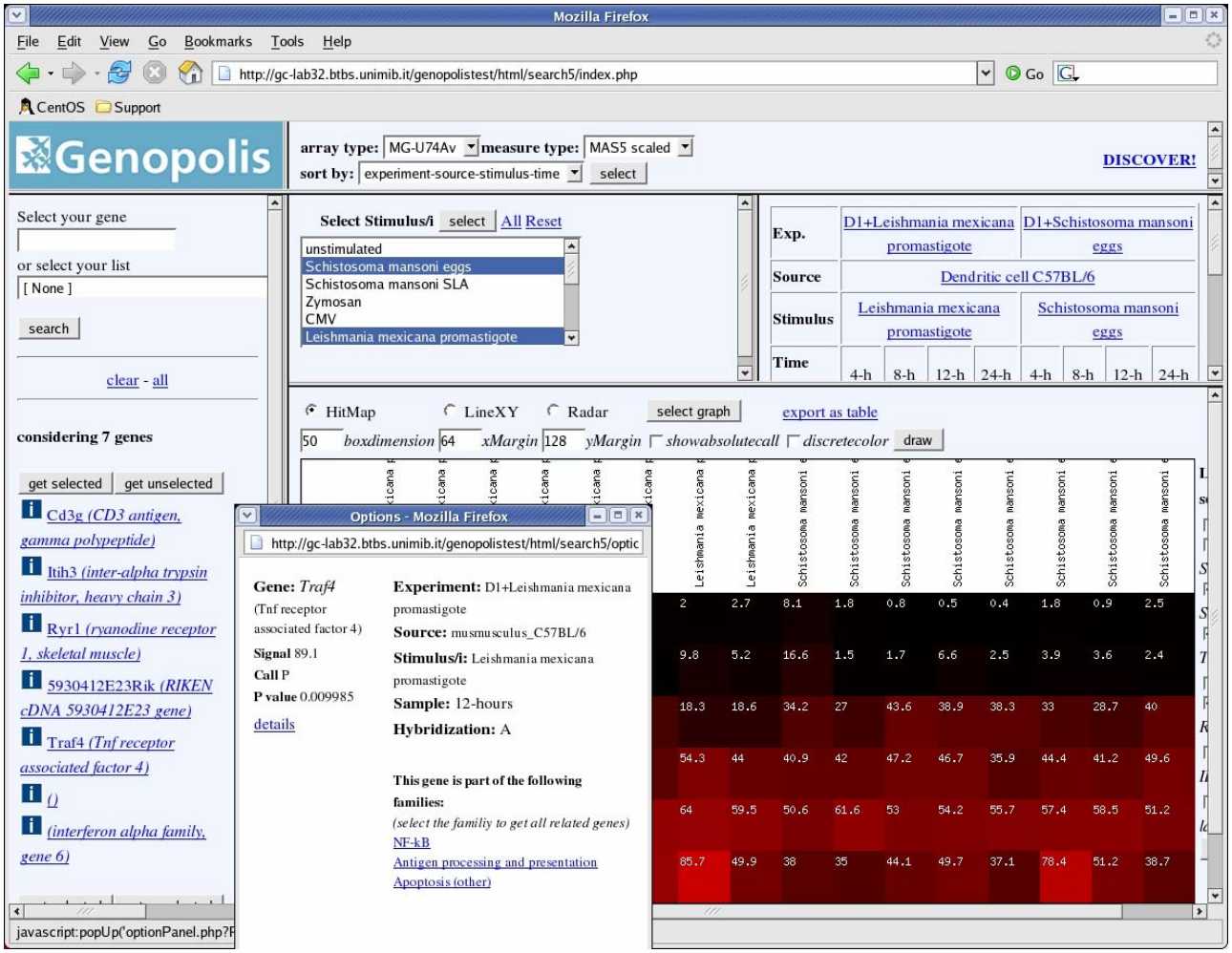
Methods
All the data have been generated through the Affymetrix platform, and all experiments are annotated following MIAME recommendations. Experiment annotation is realized through a custom software. At the core of this software is set of objects that represent entities relevant to the experiment annotation, such as Experiment, Source, Stimulus, Sample, Hybridization, Measure. These objects are organized as a tree, and the system provides functions on this tree to navigate and check its components. Permission to edit and view these objects can be defined at the object level with a group/role system of authorization. Furthermore the system supports the creation of controlled vocabularies by the users. Integrity and consistency of data and annotation is enforced through checking procedures. These are both automatic, on file integrity and required fields, and humanly supervised, as for controlled vocabularies definition. The system is based on a web architecture and is implemented in PHP and Java. In its current version it is based on MySQL and is deployed on a Linux/Apache redundant server with high availability features. Several kind of data are managed by the system. Raw images and cell files are managed as files, and made available for download to authorized users, while expression values and experiment descriptions are managed by the SQL engine and used for basic data analysis and advanced visualization. The system also offers an automated export to ArrayExpress, parsing of Affymetrix MAGE-ML description files and and advanced interactive user interface. These interface allows users to visualize data matrices based on functional lists and sample characterization, and to navigate to other data matrices defined by similarity of expression values as well as functional characterizations of genes involved. A collaborative environment is also provided for the definition and sharing of annotation by users (eg.: functional annotations of genes).

Results
The Genopolis database system provides an advanced resource for a scientific community investigating a common topic through microarrays . Roles can be defined so that measurements and experiment descriptions can be provided by different users. Consistency of data and annotations in enhanced by the system and common controlled vocabularies are built as the result of the interaction of users. Data can be kept confidential among groups for a limited time, and can be later made public to other users and on the ArrayExpress public repository. The content of the knowledge base can be exported for analysis with external tools, or browsed with an interactive graphical system that intuitively allows users to browse related set of genes and experimental conditions.

Availability: http://www.genopolis.it/

Contact email: andrea.splendiani@unimib.it

Supplementary informations
Access to data is subordinated to proper agreement.



Quantifying the relevance of different mediators in the human immune cell network

Tieri P (1,2), Valensin S (1,2), Latora V (3), Castellani GC (2), Marchiori M (4,5), Remondini D (2), Franceschi C (1,2,6)

(1) Dipartimento di Patologia Sperimentale, Università di Bologna, Bologna
(2) C.I.G.-Centro Interdipartimentale 'L. Galvani', Università di Bologna, Bologna
(3) Dipartimento di Fisica ed Astronomia & INFN, Università di Catania, Catania
(4) W3C MIT Lab for Computer Science, Cambridge, USA
(5) Dipartimento di Informatica, Università di Venezia, Venezia
(6) INRCA -Istituto Nazionale di Ricovero e Cura Anziani, Ancona

Motivation

The cells of the Immune System (IS) communicate by direct surface contact and indirectly, by means of soluble mediator proteins released and bound by the immune cells. Soluble mediators implement cellular communication both at short range and across the major body systems. The network we consider is constituted by various immune cell types, which can act as both sources and targets of the exchanged mediators. Mediators are characterized by pleiotropy (each mediator has multiple targets) and redundancy (each mediator is produced by several sources), two characteristics that strongly influence the reliability, the robustness and the adaptability of the IS. In this view we built a network of IS cells whose cell-cell interactions are mediated by soluble molecules such as cytokines, chemokines, hormones. From experimental immunological knowledge we have two types of information: 1) each cell type secretes a defined set of mediators; 2) each mediator affects a defined set of cells. Combining information from these two datasets we can write down a complete"relationship matrix" among immune cell types where the relations are constituted by exchanged soluble mediators. Following this approach we retrieved all available literature data from the online Cytokine Reference Database, choosing a set of 19 cell types involved in the most relevant immune processes and the related secreted and affecting mediators (90 proteins).

Methods

The immune cell network is represented as a valued directed graph, cell types are the vertices of the graph and the soluble mediators form its arcs: a directed arc from vertex i to vertex j is defined by the existence of at least one mediator secreted by cell i and affecting cell j. Cell self-stimulation by soluble mediators (autocriny) is also taken into account. The value eij referred to each arc is equal to the number of different mediators connecting the cell i to the cell j. We consider such a number as a measure of the importance of the communication between two cells along the arc, hence modelling this as an efficiency that measures the bandwidth. Two cells/vertices in the graph can communicate through various paths, connecting them with different levels of efficiency. We assume that the communication between vertices i and j takes the most efficient path, the one that assures the widest communication band possible. We characterize the system global properties by defining the network efficiency given by the sum of the most efficient paths that link up every couple of nodes divided by the squared number of the existing nodes. Here we propose a method for quantifying the centrality of the various soluble mediators in the IS. The method is based on the concept of efficient communication over the immune cell network. The centrality of each mediator is measured by its network relevance, defined as the relative drop in the network efficiency caused by the removal of the mediator. In fact, in our framework, the removal of a mediator weakens some of the values eij relative to the arcs and, consequently, affects the communication between various couples of cells, influencing the efficiency of some paths and thus the whole IS network efficiency.

Results

The graph has 19 vertices and 316 arcs that include self-connections, out of the 361 possible arcs, and thus a pure topological analysis would have been poorly significant. Therefore, we performed a more refined analysis by taking into account the strength of the interactions among the IS cells. The

integer values eij attached to the arcs range from 1 to 36, since there are up to 36 different mediators connecting a couple of cells. Three mediators out of 90 only, TGF-b, MIP-1-alpha and -beta, and TNF-a, show a network relevance larger than 0.5; 11 mediators have network relevance in the range [0.2, 0.5]; and, the remaining 76 have network relevance in the range [0, 0.2]. The sum of the network relevance of the first three mediators accounts for the 20.5% of total mediators relevance, while those of the second and third groups account for 27.6% and 51.9%, respectively. The three most important mediators are pro- and anti-inflammatory molecules, and are involved in the communication among a large number of cell types, i.e. in 216, 224 and 120 interactions, respectively. The second group of mediators includes 9 pro- and anti-inflammatory cytokines/chemokines, one non-inflammatory cytokine (IL-7), and one neuro-endocrine hormone (VIP/PACAP). So, notwithstanding the fact that mediators involved in the inflammatory process account only for 24% of all mediators, 86% of the top molecules are inflammatory, accounting for 50% of all inflammatory mediators. In conclusion, mediators involved in innate immunity -the most ancestral branch of the immune system- and in highly conserved defence pathways such as inflammation, appears to give a substantial contribution to the efficient communication of the IS network.

Availability: http://www.immunologia.unibo.it/en/models/models.htm

Contact email: p.tieri@unibo.it

GAIA: Generation of alternative alignments by an inverse approach

Tosatto SCE, Albiero A

Dept. of Biology and CRIBI Biotechnology Centre, University of Padova, Padova

Motivation
State-of-the-art methods for protein structure prediction based on the modelling of an unknown structure from a known template have been recently shown to achieve greater accuracy with the simultaneous modelling of ensembles of slightly different models, which have to be subsequently ranked using model quality assessment programs (MQAPs). One way to construct large ensembles of alternative models is from alternative alignments, generated either from varying alignment parameters or using different alignment methods. These approaches however do not guarantee a systematic and uniform coverage of conformational space and are therefore not guaranteed to find the most accurate solution, while spending much time exploring redundant solutions.

Methods
Here we propose a novel combinatorial optimization method for the systematic exploration of the alignment space defined by a limited set of initial alternatives. The method uses the initial alternatives to construct an inverse alignment matrix which can be traversed with regular dynamic programming to produce new solutions from suboptimal alignments. A divide and conquer step ensures the generation of truly different solutions ranked by relative frequency in all regions of alternative alignment.

Results
The inverse alignment approach was benchmarked on the comparative modelling and fold recognition homologous targets from the recent CASP-6 blind test and on a test set of 20 difficult alignments with little sequence identity. The results indicate that the inverse alignment approach is capable of ensembles of up to thousands of alternative alignments in minutes of computer time on a single desktop processor. The generated ensemble is enriched in accurate solutions and facilitates the MQAP selection of near-native models.

Contact email: silvio@cribi.unipd.it

Local protein structure validation revisited

Tosatto SCE (1), Battistutta R (2), Albiero A (1)

(1) Dept. of Biology and CRIBI Biotechnology Centre, University of Padova, Padova
(2) Dept. of Chemical Sciences and Venetian Institute of Molecular Medicine (VIMM), University of Padova, Padova

Motivation
Structure validation by computational methods is an important tool in protein crystallography. Stereochemical criteria are mainly used to distinguish locally between distorted and adequately refined models. However, the readily available criteria are not sufficient to clearly establish the global quality, producing only rough indications instead.

Methods
A new criterion, called TAP, measuring local sequence to structure fitness based on torsion angle propensities normalized against the global optimum is introduced.

Results
It is shown to correlate well with experimental parameters for global X-ray accuracy. The TAP score can be used directly to validate the correctness of refined X-ray models. It is shown to have a two to five times higher correlation with experimental quality measures than previous methods. Highly selective TAP thresholds are derived to recognize over 99% of the top experimental structures in the absence of experimental information. Estimating the local interactions can help in the accurate determination of protein conformation by experimental and computational means.

Contact email: silvio@cribi.unipd.it

On the experimental annotation of tomato BACs sequences: reliable alignment for useful genomic analyses

Traini A (1,2), Chiusano ML (2)

(1) PhD fellow in Computational Biology, Interdepartmental Research Center for Computational and Biotechnological Sciences, Second University of Naples, Naples, Italy
(2) Department of Structural and Functional Biology, University "Federico II", Naples, Italy

Motivation
Curated gene annotation is a challenging computational problem in genomics. Reliable results are commonly achieved with spliced alignment of full-length cDNAs or expressed sequence tags (ESTs) with sufficient overlap to cover the entire mRNA. Moreover, predictive approaches are based on curated Gene Models obtained by experimental effort too. Many standalone programs are available for mapping and aligning expressed tags to a genome sequence. With the aim to contribute to the bioinformatics efforts of the International Tomato Genome Sequencing project we tested the software proposed by the International Committee for cDNA/EST to genome mapping. We investigated which algorithm is more reliable and in which context, comparing and evaluating some of the most frequently used specialized software to provide a trusty Reference dataset of Tomato Gene Models.

Methods
We tested sim4 [1], based on the BLASTZ algorithm [2], Galahad, included in the grail-exp package [3], blEST [1] and SIBsim4 [1], ALL based on sim4 algorithm [1], and GeneSeqer [4][5] as software commonly used for solving the task of mapping cDNAs/ESTs to genomic sequences. Software results depend on the specific parameter usage but also on specific similarity thresholds. We compared the resulting data from different software considering all common results and discussing each software specific feature. To provide a reliable informative benchwork for Tomato genome annotation based on experimental results, we set up a Gbrowse [6] based platform reporting the results from the different methods. The datasets used in the present analysis are the Tomato expressed sequences available from dbEST [7] and from the genome sequencing effort at the SOL Genomics Network (SGN)[8].

Results
This work summarizes the analysis and the results of specific software solving the cDNA/EST to genome mapping problem. Different algorithms and even different parameters and similarity thresholds influence the quality of the resulting alignments. We present here our evaluation of the software considered and we propose multiple algorithm usage under different constraints to provide exhaustive and reliable information when experimentally annotating genome sequence data. indeed To further support annotators, we provide different results: i) best alignments with a low error margin of the genomic region alignment and ii) and less stringent results extending the number of retained alignments. This may be useful because highly scored complete EST alignment as well as medium level similarities and partial alignments per EST may be of interest, either when considering reliable gene models predicted by experimental data or when looking for related gene loci in evolutionary analysis.

Availability: http://www.cab.unina.it/

Contact email: chiusano@unina.it

References
1. Florea L, Hartzell G, Zhang Z, Rubin GM, Miller W. A computer program for aligning a cDNA sequence with a genomic DNA sequence. Genome Res. Sep;8(9):967-74 (1998).

2. S. Schwartz, W. J. Kent, A. Smit, Z. Zhang, R. Baertsch, R. C. Hardison, D. Haussler and W. Miller, Human-Mouse Alignments with BLASTZ, Genome Res. Vol 13, Issue 1, 103-107, January (2003).
3. D. Hyatt, J. Snoddy, D. Schmoyer, G. Chen, K. Fischer, M. Parang, I. Vokler, S. Petrov, P. Locascio, V. Olman, Miriam Land, M. Shah, and E. Uberbacher, Improved Analysis and Annotation Tools for Whole-Genome Computational Annotation and Analysis: GRAIL-EXP Genome Analysis Toolkit and Related Analysis Tools, Genome Sequencing & Biology Meeting, May 2000
4. Brendel V, Xing L, Zhu W. Gene structure prediction from consensus spliced alignment of multiple ESTs matching the same genomic locus. Bioinformatics. May 1;20(7):1157-69. Epub 2004 Feb 5 (2004).
5. Shannon D. Schlueter, Qunfeng Dong and Volker Brendel, GeneSeqer@PlantGDB: gene structure prediction in plant genomes, Nucleic Acids Research, Vol. 31, No. 13, 3597-3600, (2003).
6. Stein, L.D., Mungall, C., Shu, S., Caudy, M., Mangone, M., Day, A., Nickerson, E. Stajich, J.E., Harris, T.W., Arva, A. and Lewis S. (2002) The generic genome browser: A building block for a model organism system database. Genome Res. 12, 1599-1610.
7. Boguski M.S., Lowe T.M., Tolstoshev C.M.,dbEST-database for "expressed sequence tags", Nat Genet., Aug; 4(4): 332-333 (1993). home-page: http://ncbi.nih.gov/dbEST/
8. home-page: www.sgn.cornell.edu

Supplementary informations

This work is supported by the Agronanotech Project (MIPAF, ITALY)

Data handling strategies for high throughput pyrosequencers

Trombetti GA (1,2), Bonnal RJP (2), Rizzi E (2), De Bellis G (2), Milanesi L (2)

(1) Consorzio Interuniversitario Lombardo per l'Elaborazione Automatica, Milano
(2) Istituto di Tecnologie Biomediche, Consiglio Nazionale delle Ricerche, Milano

Motivation

New high troughput pyrosequencers such as the 454 Life Sciences GS 20 are able to massively parallalize DNA sequencing providing an unprecedented rate of output data and potentially reducing costs. However, these new pyrosequencers also bear a different error profile and provide shorter reads than those of a more traditional Sanger sequencer. These facts pose new challenges regarding how the data are handled and analyzed. More in detail: - Different error profile: new high throughput pyrosequencers provide good performances on average, however, these sequencers have problems in correctly basecalling on homopolymers. - Shorter reads: pyrosequencers nowdays only provide very short reads of about 94 bases on average. This poses problems when a reference sequence is not available or when detecting mutations in repeats or low complexity areas. - Rate: a running 454 LifeSciences GS 20 pyrosequencer can analyze up to 10MBases/hour or 110,000 90-bases-reads/hour. The computation system and data handling strategy should be able to accommodate this.

Methods

To address the challenges described in the above paragraph, we created an automated calculation pipeline integrated with a database storage system. The database is capable of storing, indexing and handling the following information: - Multiple projects of analysis - Biological samples and protocols - Sequences read by the GS 20 sequencer for each run - Multiple co-existing databases of reference sequences - Final results of the calculation pipeline, such as punctual mutations found (with support for heterozygosity) - Intermediate calculations of the calculation pipeline, such as Blast results The pipeline together with the database storage system is capable of easily repeating any past computation thus demonstrating any results obtained, in addition, it is possible to repeat the computations with altered parameters, constants and thresholds and easily compare the results leveraging the database functionalities. The database also allows our biologists to perform inspired researches starting from what they see from the "stock" results obtained from the pipeline. This allows us to discover and investigate peculiar phenomena more easily. The pipeline is multi-stage and mostly parallelizable. Here we quickly outline how we addressed the challenges mentioned in the motivation: Different error profile: Attention is paid in the algorithm so that a nearby homopolymer does not trigger false punctual mutations. Raising the coverage multiplicity only helps up to a point on kinds of errors happening consistently. Shorter reads: The speed and lower operation costs of the 454 machine allowed us to use a high coverage, so that most sporadic errors fall under a threshold. In case of regions similar on more than one reference sequence an heuristic algorithm discriminates which matches are to be kept and which are to be discarded. High data rate: The parallelizability of the pipeline ensures it can keep up with the high data rate of the 454 machine as well as the possibility that biologists would want to repeat a large number of past calculations with different parameters. The database storage system has been set up to store years of operation in an organized and searchable fashion. Backups are made via incremental diffs.

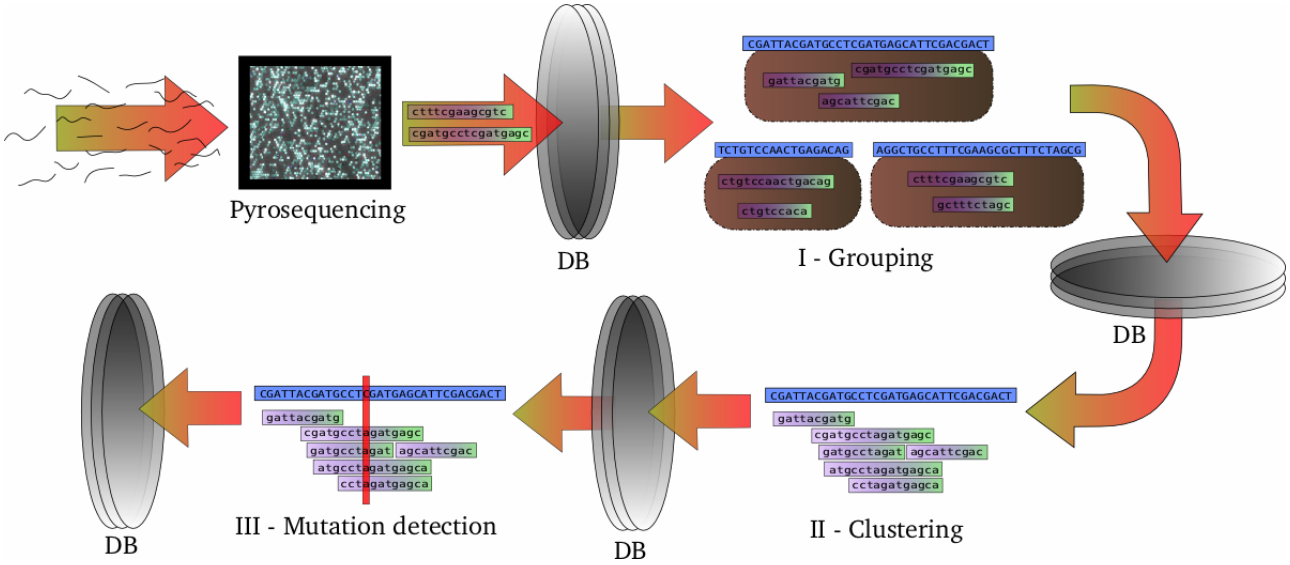
Results

After various runs of the pipeline aimed at tuning the parameters and thresholds for optimal results, we were able to successfully analyze 273 sequenced amplicons from chromosomes 1, 3, 5, 9, 11, 13, 17, 18, 19 of a human sample and correctly find punctual mutations confirmed by either NCBI dbSNP or Sanger resequencing. Sequencing was made with our 454 Life Sciences GS 20 pyrosequencer, obtaining 500,000 reads of 94 bases on average. More analyses will be performed in the future. This pipline is realised in the frame of the Italian MIUR-FIRB project LITBIO www.litbio.org (Laboratory for Interdisciplinary Technologies in Bioinformatics).

Contact email: gabriele.trombetti@itb.cnr.it

Supplementary informations

Trombetti G. A. is a Ph.D student from DEIS department - University of Bologna, Italy



Structural adaptation to low temperatures: analysis of the subunit interface of oligomeric psychrophilic enzymes

Tronelli D, Gianese G, Pascarella S

Dipartimento di Scienze Biochimiche "A. Rossi Fanelli" Università "La Sapienza", 00185 Roma, Italy

Motivation

Psychrophiles are ectothermic organisms adapted to life in cold permanent environment. Enzymes from such organisms show a higher catalytic efficiency in the 0 - 20 °C temperature range when compared to mesophile, thermophile and hyperthermophile homologues. This is usually associated to a lower thermostability. Physical and chemical characterization of these enzymes is currently under study in order to understand the molecular basis of cold adaptation. Psychrophilic enzymes are often characterized by a higher flexibility which allows for a better interaction with substrates, and by lower activation energy requirement if compared to mesophile and thermophile counterparts. In their tertiary structure, psychrophilic enzymes present fewer stabilizing interactions, longer and more hydrophilic loops, higher glycine and lower proline and arginine content. Protein surfaces often show a lower charged amino acid content and a high number of hydrophobic side-chains.

Methods

In this study, we carry out a comparative analysis of the structural characteristics of the interfaces between oligomeric psychrophilic enzyme subunits. Crystallographic structures of oligomeric psychrophilic enzymes, their mesophile homologues (and, when available, also thermophile and hyperthermophile enzymes) belonging to five different protein families were retrieved from Protein Data Bank. The following structural parameters were calculated from the atomic coordinates of each enzyme within its family: overall and core interface area, characterization of polar and apolar contributes to the interface, ion pair number and hydrogen bonds between monomers, internal area and total volume of non solvent-exposed cavities at interface, average packing of interface residues. These properties were compared to those of mesophile, thermophile and hyperthermophile enzymes. Results were analysed using Student's unpaired two-tailed t-test.

Results

The comparative analysis reveals that some of the differences observed within each family could be attributed to cold temperature adaptation. The most significant differences between psychrophilic and mesophilic proteins are found in the number of ion pairs and the number of hydrogen bonds. Psychrophilic proteins, moreover, show a significant decrease in the apolarity of their subunits interface.

Contact email: daniele.tronelli@uniroma1.it
stefano.pascarella@uniroma1.it

Genes clustering on large, mixed microarray data sets

Tulipano A (1,2), Marangi C (4), Angelini L (3), Pellicoro M (3), Donvito G (2), Maggi G (2), Gisel A (1)

(1) CNR, Istituto Tecnologie Biomediche Sezione di Bari, via Amendola 122/D, 70126 Bari (Italy)
(2) INFN Sezione di Bari, via Amendola 173, 70126 Bari (Italy)
(3) Dipartimento Interateneo di Fisica, Università di Bari, via Amendola 173, 70126 Bari (Italy)
(4) CNR, Istituto per le Applicazioni del Calcolo Sezione di Bari, via Amendola 122/D, 70126 Bari (Italy)

Motivation
Every single microarray data set is an image of the transcriptional level of ten thousands genes during a specific biological experiment at a specific time. An increase in the number of data sets within the biological experiment multiplies the number of images and forms an overall picture of the processes monitored during the biological experiment. Every picture contains information of some general processes found also in other biological experiments and some information on some processes which are specific for that biological experiment. Public databases such as GEO (http://ncbi.nlm.nih.gov/geo) and ArrayExpress (http://www.ebi.ac.uk/arrayexpress) contain a large amount of such biological experiments including several data sets. Those large numbers of data sets can be used to monitor and differentiate general processes and specific processes within such biological experiments.

Methods
As a first test we were downloading from GEO a data collection derived from the Affimetrix microarray design 'Human Genome U133 Array Set HG-U133A' all normalized by MAS 5.0. The total number of data sets we included in our analysis is 587 covering more than 20 biological experiments. In order to have a comparable set of data, we scaled each data set point by means of a global normalization, doing a logarithmic transformation on it and setting the median of the values distribution of each microarray experiment to zero. For the global clustering analysis we have chosen a clustering algorithm based on the cooperative behaviour of an inhomogeneous lattice of coupled chaotic maps, the Chaotic Map Clustering (CMC, Angelini et al. 2000). A chaotic map is assigned to each data point and the strength of the coupling between pairs of maps is a decreasing function of their distance. The mutual information between pairs of maps, in the stationary regime, is then used as the similarity index for clustering the data set. For our analysis we set the parameter for the cluster resolution high enough to observe clusters as stable as possible. Running CMC under such stringent conditions on the whole set of 587 microarray experiments generated few clusters containing no more than 40 genes per cluster.

Results
From the analysis of the members of each cluster by the Gene Ontology (http://www.geneontology.org/) it is clear that those clusters contain genes representing biological processes very general, such as metabolism of different compounds, different transport processes, RNA processing, but also clearly different among each other. Limiting now the analysis to a subset of biologically similar experiments, we find the same clusters as in the global analysis in addition to some new and therefore specific clusters for that subset of microarray experiments. In this way, by choosing different subsets of microarray experiments, we are able to assign to each subset specific biological processes and find new annotations for genes little annotated within those clusters. Such an analysis takes advantage of the large information within those high throughput data publicly available to improve the knowledge of every single gene represented in those data set.

Contact email: angelica.tulipano@ba.infn.it

References

- Angelini L., De Carlo F., Marangi C., Pellicoro M. and Stramaglia S., 2000 Clustering data by inhomogeneous chaotic map lattices. Phys. Rev. Letters 85(3); 554-557.

False occurrences of functional motifs on protein sequences highlight evolutionary constraints

Via A (1), Gherardini F (1), Ferraro E (1), Scalia Tomba G (2),
Ausiello G (1), Helmer-Citterich M (1)

(1) Centro di Bioinformatica Molecolare, Department of Biology, University of Tor Vergata, Roma
(2) Department of Mathematics, University of Rome Tor Vergata, Roma

Motivation
A functional motif is a set of residues that is characteristic of a specific biochemical function. The detection of a functional motif in yet uncharacterized protein sequences is a well-established method for assigning function to proteins. A critical problem, however, concerns the evaluation of the false prediction rate of a motif in sequence databases, i.e. the significance of finding a motif in several proteins. The number of false positive (FP) matches of a pattern has been often assessed from the number of its occurrences expected by chance (E) for the mere aggregation of letters in a database search, as can be calculated from the residues frequency in the database ([1], [2], [3], [4]). The relationship between E (expected) and FP (observed), however, has not been thoroughly investigated so far. It is reasonable to expect that the function fitting the set of data (E,FP) is linear, but it is not clear a priori if there are exceptions (i.e. number of false predictions on a biological database sensitively greater or lower than the expected number of hits on the corresponding random database), how frequent they are, and the reason why they occur.

Methods
In this work, we carried out a statistical study of such relationship and an analysis of the unexpected behaviours, thus providing insights into the random nature of protein sequences. The analysis described in this work was performed on 1226 PROSITE patterns in the form of regular expression and based on three sequence datasets: the complete Swiss-Prot database (sprot100), the set of H. sapiens sequences, and the set of S. cerevisiae sequences derived from sprot100. We assumed for the expectation E of a pattern P, the mean number of hits on N database randomizations. In order to preserve the local sequence composition the datasets were randomized by reshuffling each single sequence. As a control, the statistical analysis was also performed on PROSITE reversed patterns, which represent a reliable sample of non-functional patterns.

Results
Our results show that a) the relationship (E, FP) is linear, b) the great majority of functional motifs (group II) have a number of false occurrences comparable to the number of matches on a random database, c) there is a group (group I) of PROSITE patterns for which FP >> E and another one (group III) for which E >> FP. Both groups I and II are outside the 95% confidence interval around the value E = FP. A detailed analysis of patterns belonging to each group revealed several interesting features. In particular patterns belonging to different groups do share specific statistical and biological properties. We compared groups using the patterns information content ([1], [2]) - as a statistical parameter - and the tendency of their false positive hits of being in either disordered or ordered/globular regions of proteins, as a biological parameter. Our findings suggest diverse fascinating mechanisms and constraints occurring during evolution, which might "regulate" the random appearance of functional motifs in protein sequences.

Contact email: allegra@cbm.bio.uniroma2.it

References
1. Jonassen, I., Collins, J.F. & Higgins, D.G. (1995) Finding flexible patterns in unaligned protein sequences. Protein Sci. 4, 1587-95.

2. Jonassen, I., Eidhammer, I., Grindhaug, S.H. & Taylor, W.R. (2000) Searching the protein structure databank with weak sequence patterns and structural constraints. J.Mol.Biol. 304, 599-619.
3. Nevill-Manning, C.G., Wu, T.D. & Brutlag, D.L. (1998) Highly specific protein sequence motifs for genome analysis. Proc. Nat. Acad. Sci. USA 95, 5865-5871
4. Sternberg, M.J.E. (1991) Library of common protein motifs. Nature 349, 111.

Valuation study of available genomic data storage platforms

Viti F (1), Merelli I (1), Porro I (2), Papadimitropoulos A (2), Milanesi L (1)

(1) Institute for Biomedical Technologies, National Research Council, Milan
(2) Department of Computer Science, Control Systems and Telecommunications, University of Genoa, Genoa

Motivation

One of the most important problems in microarray research is the storage of large amounts of data, keeping track of the information about experiments, samples and projects in matter. Nowadays, there are three recognized public repositories for microarray experiments: Gene Expression Omnibus (GEO - NCBI), ArrayExpress (EBI) and the Center for Information Biology Gene Expression Database (CIBEX - DDBJ). These infrastructures can be used to deal with high-throughput experimental data in gene expression research, and they are all MIAME (Minimum Information About Microarray Experiment) compliant. Scientists who work with microarray data need more than an infrastructure for data sharing: the technique is difficult to handle not only because it produces thousands of data, but also because it requires many biological and technological steps that must be recorded. So they need a secure local storage system to manage and integrate broad genomic data, in order to freely insert, change and modify designs and protocols of their experiments, always following strict standards approved by all of the microarray community. In this context our aim is to test infrastructures the bioinformatics world proposes and to evaluate the most efficient genomic platform available at the moment. Considering that some important infrastructures (i.e. GEO and CIBEX) do not have a local downloadable version, we analysed ArrayExpress and compared it with another important platform, maybe less known but with great potentialities: GUS, the Genomics Unified Schema. It is an integrated databases system, developed by the enormous contribution of University of Pennsylvania, non completely avaliable as web service but on which schema of important on line databases like RAD (RNA Abundance Database) are based.

Methods

To make an evaluation of the capabilities of the data warehousing of these open source products we installed both of them in order to test their features in our microarray data and our experimental protocols. The two platforms could seem comparable, but analyzing them in detail some important differences emerge. GUS shows interesting aspects about user interface and scheme customization, which are not so rigid as in ArrayExpress. Moreover, its data integration in genomics, transcriptomics and proteomics fields presents good features for future bioinformatics studies. Concerning requirements, they are similar on various aspects: they both need a Unix Operative System; both suggested Oracle as RDBMS, even if, while ArrayExpress is a combination of two different databases (i.e. Oracle and MySQL) to separate data storage from query data warehouse, GUS uses a unique database. Loading proper data or information from external databases is different in the two systems. GUS perform this process by using and creating new Perl plug-ins (which generate Perl objects), while ArrayExpress provides the MAGEloader/MAGEvalidator, a java program which converts data into MAGE-ML (Microarray Gene Expression Markup Language), language derived from MAGE-OM (MicroArray Gene Expression Object Model) and similar to XML. MAGE-OM is, in fact, ArrayExpress fixed logic schema, while GUS as a proper schema that can be modified and also enlarged by single developers. Regarding data submission standards, both are MIAME-compliant, and GUS follows developed standards also for proteomics (MAIPE - Minimum Information About a Proteomics Experiment) and tissue gene expression localization (MISFISHIE - Minimum Information Specification For In Situ Hybridization and Immunohistochemistry Experiments) experiments. Both platforms can be deployed, through the Apache Tomcat servlet container, and visualized as web applications: an example of this possibility in GUS is RAD website, which makes accessible a part of the integrated databases system of GUS. Even if these two infrastructures seem to be quite similar, there are also important differences, particularly inherent database schema and UI structure. Query and data submission in ArrayExpress

are pre-configured and cannot be modified, while GUS platform offers the possibility to manage, through a Web Development Kit, the user interface, with the aim of creating user-friendly applications and websites with advanced query capabilities. This is not a secondary aspect because bioinformatics databases are often handled by biologists who need a simple, linear schema to insert and make queries on information.

Results

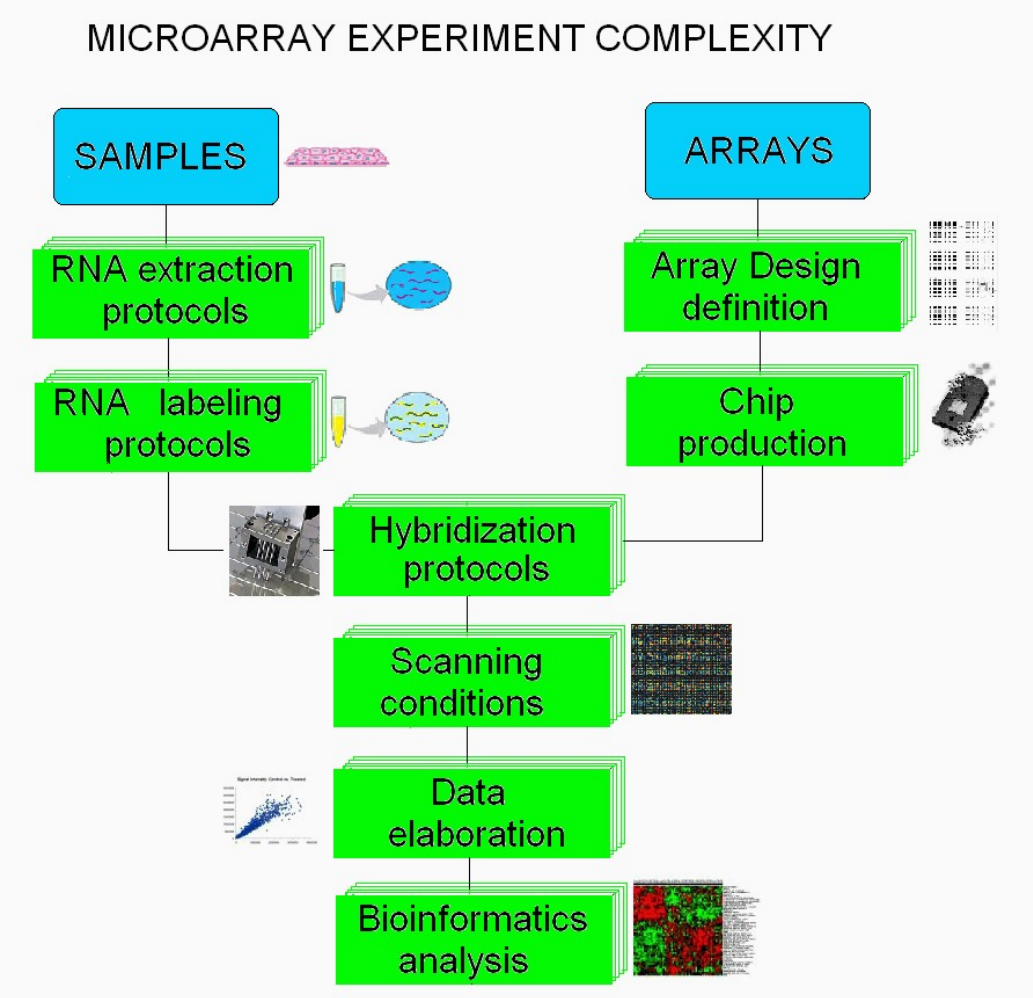
In order to valuate a powerful platform to handle biologically consistent results, we noticed that GUS is highly customizable in structure and flexible in the user interface development. Moreover, ArrayExpress can store data from all microarray technologies and from array-based chromatin immunoprecipitation and array CGH, while GUS permits submission of different kinds of data, from genomics, to transcriptomics and proteomics, in an aim to improve and support broad genomics data integration.

Availability: http://www.gusdb.org/

Contact email: federica.viti@itb.cnr.it

Supplementary informations

GUS is available at http://www.gusdb.org/
ArrayExpress is available at http://www.ebi.ac.uk/arrayexpress/
GEO is available at http://www.ncbi.nlm.nih.gov/geo/
CIBEX is available at http://cibex.nig.ac.jp/index.jsp



A Global Gene Evolution Analysis on Vibrionaceae Family Using Phylogenetic Profile

Vitulo N (1),Vezzi A (1), Campanaro S (1), Romualdi C (1), Lauro F (2), Valle G (1)

(1) CRIBI biotechnology centre, Department of Biology, University of Padova
(2) SCRIPPS institution of oceanography, University of California, San Diego

Motivation

In the past ten years, thanks to the technology advance, a great number of microbial genomes have been sequenced and a huge amount of genes have been stored in the public databases. This gave the possibility to identify the gene core shared by all the organisms, the genes characteristic of particular group of microbes and laterally transferred ORFs (alien DNA). The phylogenetic profile is based on the observation that genes involved in the same metabolism or structural complex tend to be both present or absent within genomes (Pellegrini et al. 1999). This allowed the prediction of hypothetical proteins function on the basis of phylogenetic profile shared with known genes. Moreover this approach is useful to identify cluster of genes shared by bacteria that are not phylogenetically related, suggesting possible laterally transferred elements. The Vibrionaceae family represents a significant portion of the culturable eterotrophic bacteria of the sea; they strongly influence nutrient cycling and various species are also devastating pathogens. In this work we propose a phylogenetic profile analysis performed on Vibrionaceae sequenced genomes using a gene distance calculation method based on substitution matrix of all orthologous genes. We applied this approach to study the evolution of the Vibrionaceae family on the basis of the gene content, identifying genes that are specific to the Vibrionaceae family and genes that are laterally transferred.

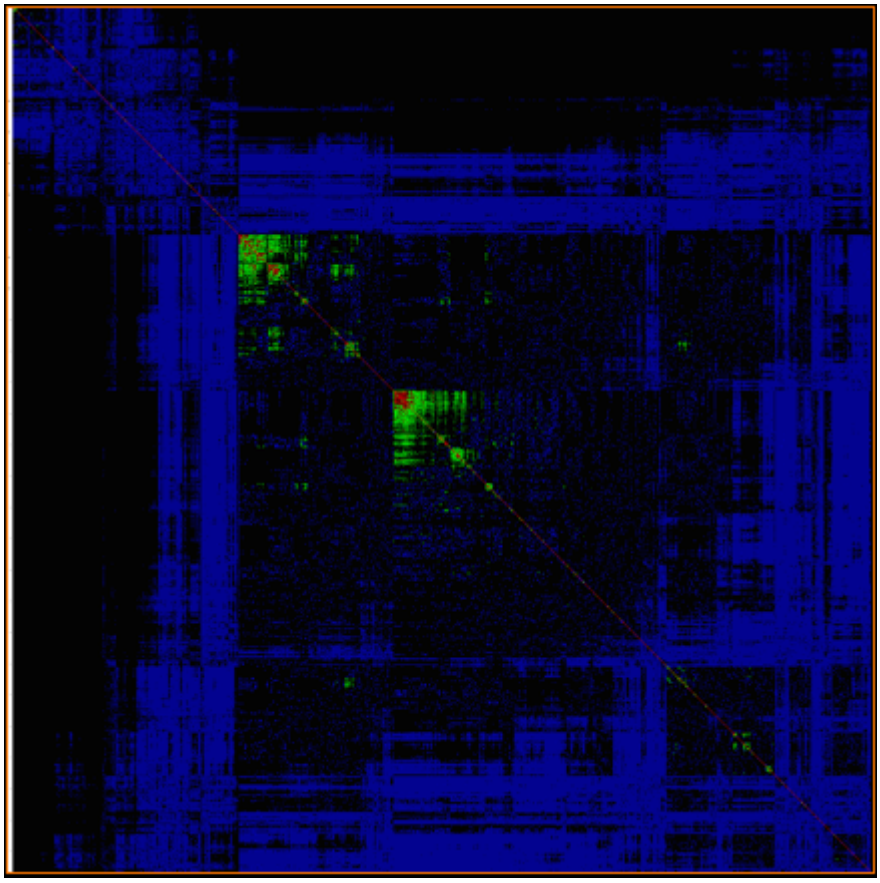
Methods

Firstly we recovered all the predicted ORFs of the completed Vibrionaceae genomes, generating a redundant list. In order to reduce the redundancy we grouped all the ORFs according to COG annotation. The remaining proteins not related to any COG entry were clustered based on similarity search. We aligned all the ORFs against all the available bacterial proteomes, using BLAST. For each ORF we generated a phylogenetic profile encoded by an array of values representing the distance calculated as the number of amino acid substitutions between the gene and the orthologue, weighted using a substitution matrix. With this phylogenetic profile we produced a distance matrix of the clusters where each element is the median distance value of the ORFs belonging to each cluster. The matrix generated underwent a cluster and a pairwise correlation analysis that produced groups of ORFs sharing similar phylogenetic profiles (see figure attached).

Results

Results shown in the figure were obtained using Photobacterium profundum as a reference and calculating the distances from the orthologous ORFs from other bacteria in order to generate the distance matrix. Obviously the analysis was restricted to ORFs present in the reference organism, while we are now performing a study considering all the Vibrionaceae ORFs as described in the "method" section. The figure shows the pairwise correlation matrix calculated using the distance measured between phylogenetic profiles. Colors varying from red to green, blue and black represent increasing phylogenetic profile distance. With this analysis we could answer many biological question on the Vibrionaceae genomic features. First of all it is possible to suggest a function for genes annotated as unknown on the basis of their phylogenetic distance from well characterized ORFs. Another interesting aspect is the identification of laterally transferred genes which have an important role on the pathogenicity of this group of bacteria.

Contact email: nicolav@cribi.unipd.it



A bioinformatic approach for a structural analysis of protein phosphorylation sites

Zanzoni A, Gherardini F, Ausiello G, Via A, Helmer-Citterich M

Centre for Molecular Bioinformatics, Department of Biology, University of Rome Tor Vergata.

Motivation

The phosphorylation of specific protein residues is a crucial event in the regulation of several cellular processes, acting on activation, deactivation or recognition of the target protein. A great amount of eukaryotic cell proteins (30 up to 50% of the total) undergoes such post-translational modification. The recent improvement in the experimental identification of phosphoproteins and phosphoresidues has increased dramatically the amount of phosphorylation sites data and the need of computational tools for collecting and analysing this data has grown accordingly. In the past years several sequence-based methods to predict phosphorylation sites were developed using different approaches such as regular expressions with context-based rules, Position-Specific Scoring Matrices (PSSMs) and artificial neural networks. Only approximately one tenth of known kinases have known consensus sequences, which often are not present in all known in vivo substrates. The structural basis and the determinants of interaction specificity are often unclear. The presence of structural determinants that only sometimes overlap with sequence consensi and that might be independent on the residue order in protein sequences might explain the problems encountered so far in unravelling the rules of kinase specificity.

Methods

We have developed a procedure for the annotation and analysis of the three-dimensional structure of experimentally verified protein phosphorylation sites, also called instances, retrieved from the phospho.ELM database. The correspondence between phospho.ELM sequences and the PDB chains was established via the Seq2Struct resource, an exhaustive collection of annotated links between UniprotKB and PDB sequences. Links are based on sequence alignment using pre-established highly reliable thresholds. For each instance mapped onto PDB chains, a structural neighbourhood, that we call zone, was defined using a distance criterion. A procedure was implemented in order to annotate each residue belonging to the defined zones with diverse functional information such as solvent accessibility, secondary structure assignment and sequence conservation. Furthermore, we performed an all versus all comparison amongst the phosphorylation zones in order to find statistically significant local structural similarities.

Results

All this information, as well as the results of a large-scale local structural comparison with stringent parameters, was stored in a publicly available relational database called Phospho3D [http://cbm.bio.uniroma2.it/phospho3d]. Amongst the structural comparison results, we selected two potentially interesting cases: a structural match that allows the inference of which kinase could be responsible for the phosphorylation of a specific tyrosine residue and an interesting candidate 3D motif in common between two substrates. In the latter structural motif, some residues are conserved both in sequence and in structure, some other are only conserved in structure and not in sequence; an experimental test is being designed to evaluate its biological relevance.

Contact email: andreas@cbm.bio.uniroma2.it