Abstract 1
- HOMOLOGOUS GENE FAMILIES DATABASES FOR COMPARATIVE GENOMICS -

Penel Simon[1], Duret Laurent[1], Gouy Manolo[1], Perriere Guy*[1]

- [1]*Laboratoire de Biométrie et Biologie Evolutive, University of Lyon ~ Villeurbanne ~ France*

Motivation: Since the availability of a huge number of sequences, comparative genomics is a central step in many sequence analysis studies. For instance, it is used to help identify regions of interest in DNA sequences, to study evolution at the molecular level (speciation events, gene duplications, whole genome duplication, etc.), to determine phylogeny of species or to predict the function of a new gene. In that context we developed a set of three homologous gene families databases that can be used in many aspects of comparative genomics. Those databases – HOVERGEN, HOGENOM and HOMOLENS – share the same architecture, and they include protein and nucleotide sequences, alignments and phylogenetic trees.
Methods: The databases are built using protein sequences from different sources. HOVERGEN contains vertebrate sequences taken from UniProt. HOGENOM is devoted to completely sequenced organisms, and its sequences come from various sources (Genome Reviews, JGI, Ensembl, Bacterial species from the NCBI, etc.) HOMOLENS is devoted to the completely sequenced eukaryotes found in Ensembl. For the three systems, protein sequences are clustered into homologous families, and then alignments and trees are built on those families. The large-scale similarity searches required to cluster sequences, as well as the massive alignments and tree computations, are performed on a cluster containing more than 2000 CPUs. Lastly, phylogenetic trees are reconciled with a reference species tree.
Results: The three databases can be fully downloaded from the PBIL (Pôle Bioinformatique Lyonnais) site (http://pbil.univ-lyon1.fr). On-line access is also provided through three different ways: query forms on the PBIL site, a general retrieval system (Query) and a devoted client-server graphical interface (FamFetch). The later can be used to perform tree-patterns based searches allowing, among other uses, to retrieve easily set of orthologous genes thanks to phylogenetic criteria.

Abstract 2
- THE GENIUS GRID PORTAL AND THE ROBOT CERTIFICATES : A NEW TOOL FOR E-SCIENCE -

Donvito Giacinto[1], Maggi Giorgio Pietro[1], Barbera Roberto[2], La Rocca Giuseppe*[3], Milanesi Luciano[4], Falzone Alberto[5]

- [1]INFN ~ Bari ~ Italy - [2]INFN and University ~ Catania ~ Italy - [3]INFN ~ Catania ~ Italy - [4]Instituto di Tecnologie Biomediche – CNR ~ Milan ~ Italy - [5]NICE S.r.l. ~ Cortanze (AT) ~ Italy

Motivation: Grid technology, based on opens standards and protocols, is the computing model which allows users to share a wide pletora of distributed computational resources regardless of their geographical and Institutional location. Up to now, the high security policy requested to access the distributed computing resources is a rather big limiting factor to increase the usage of Grids to a wider community of users. Grid security is indeed based on the public key infrastructure of X.509 certificates and the procedure to get and manage those certificates is unfortunately not straightforward.

Methods: A notable step forward to increase the exploitation of this new paradigm, has recently been made with the adoption of robot certificates. These new certificates have been introduced to permit users to easily use Grid infrastructures for their research activity. The robot certificate, associated to the specific application a user wants to share with all the members of a given VO, can be installed on a smart card and used with a portal by everyone interested in running that application in a Grid environment using an user-friendly graphic interface. In this work, the EnginFrame framework of the GENIUS portal has been extended in order to support the new user's authentication based on the use of robot certificates stored on smart cards. When the smart card is inserted in the server where GENIUS is running, the portal will start generating a new user's proxy signed by the robot certificate, otherwise the normal authentication based on a dedicated Java applet will be performed. Once the proxy is generated the user is automatically redirected to the home page of the application associated with the certificate. Any other attempts to access to unauthorized applications will be blocked by the portal. Moreover, in order to enhance the security of the system an User Tracking System has also been introduced to register and monitor the most relevant actions performed by users.

Results: The work carried out and reported in this contribution is particular relevant for all users who are not familiar with personal digital certificates and the internals of the Grid Security Infrastructure. The valuable benefits introduced by robot certificates in e-Science can so be extended to users belonging to several scientific domains, providing an asset in raising Grid awareness to a larger number of potential users.

Abstract 3
- IN SILICO PREDICTION OF ESCAPE MUTANTS OF THE HIV-1 PROTEASE -

Agramonte Alina*[1], Pajón Rolando[3], Carrasco Ramón[4], Padrón Juan Alexander[5]

- [1]Bioinformatic Group, University of Informatic Sciences ~ Havana ~ Cuba - [3]Centre for Genetic Engineering and Biotechnology ~ Havana ~ Cuba - [4]Centre for Pharmaceutical Chemistry ~ Havana ~ Cuba - [5]Laboratory for Theoretical and Computational Chemistry, Chemistry Faculty, Havana University ~ Havana ~ Cuba

Motivation: The "reverse vaccinology" is a new approach that allows the in silico study of a vaccine candidate. This method reduces the time needed for the identification of these vaccines candidates and increases the success rate in those that conventional ways seems impossible. Its principal weakness consists of the experimental appearance of escape mutants at short-medium time. In this work, a first approach to predict the emergence of drug-resistant mutants under positive selection in the HIV-1 protease, correlating structural variables with the occurrence of viable mutants at population level, is proposed. Single-point mutants in a protein evolving under positive selection pressure don't randomly occur. They depends on the structural capability of the protein to accept changes without compromising function; that's why it is necessary to predict those sites where it is more likely to appear viable mutations which can be selected as escape mutants after drug treatments.

Methods: To carry out the analysis, several bioinformatics tools were used. These tools are integrated on a distributed calculation platform implemented in our University. All drug-resistant mutants of human immunodeficiency virus type 1 (HIV-1) protease, were retrieved from Los Alamos HIV drug resistance database (http://hiv-web.lanl.gov). Structural models were generated and the energy contribution in the stability of the protein for each single point mutant was evaluated.

Results: Maximum likelihood analysis provides strong evidence of positive selection acting on 19 residues of the HIV-1 protease. This number represents 48% of the total drug-resistant mutants reported in the database, until February 2007. Most of the analyzed drug-resistant mutants favorably contribute to the stability of the protein structure. They show a good correlation between the susceptibility to the occurrence of positive selection on a single point mutation and a favorable contribution of them to the stability of the protein structure. With these results, the computational Grid technology is put in hands of the researchers as an efficient tool to detect viable positive selective mutants. In this sense, it is possible to say that this strategy can be the first step to design a rational automated approach for the search of vaccine candidates to this specific therapeutic target and some others.

Abstract 4
- MASSIVE NON NATURAL PROTEINS STRUCTURE PREDICTION USING GRID TECHNOLOGIES -

Minervini Giovanni[1], La Rocca Giuseppe[2], Evangelista Giuseppe[1], Luisi Pier Luigi[1], Polticelli Fabio*[1]

- [1]Department of Biology, University Roma Tre ~ Rome ~ Italy - [2]INFN-Catania ~ Catania ~ Italy

Motivation: The number of natural proteins is an infinitesimal fraction of all the theoretically possible protein sequences. In fact, considering random protein sequences of only 100 amino acids it is possible to obtain 100^20 structurally different proteins. Thus, there is an enormous number of proteins never exploited by nature or, in other words, “never born proteins” (NBPs). A fundamental question in this regard is if the ensemble of natural proteins possesses peculiar properties in terms for example of thermodynamic, kinetic or functional properties. A key feature of natural proteins is the ability to form a stable and well defined three-dimensional structure. Thus, the structural study of NBPs can help to understand if natural protein sequences were selected during molecular evolution for their peculiar properties or if they are just the product of contingency. This problem cannot be approached experimentally, as this would require the structural characterization of a huge number of random proteins. Thus we chose to tackle the problem using a computational approach.
Methods: A random protein sequences library (2x10^4 sequences) was generated using the utility RandomBlast which produces random amino acid sequences with no significant similarity to natural proteins. The structural properties of NBPs were studied using the ab initio protein structure prediction software Rosetta (Rohl et al. Methods Enzymol. 2004; 383, 66-93). Given the highly computational demanding problem, the Rosetta software was ported in the EUChinaGRID infrastructure (http://www.euchinagrid.org) and a user friendly job submission environment was developed within the GENIUS Grid Portal (https://genius.ct.infn.it/). Protein structures generated were analysed in terms of secondary structure content, overall topology, surface/volume ratio, hydrophobic core composition, net charge.
Results: Results obtained indicate that the vast majority of NBPs, according to the Rosetta model, are characterized by a compact three-dimensional structure with a high secondary structure content. Structure compactness is comparable to that of natural proteins, suggesting similar stability. Deviations are observed in hydrophobic core composition, as NBPs appear to be richer in aromatic amino acids with respect to natural proteins. The results will be discussed in view of the evolutionary implications of NBPs properties both at the amino acid and nucleotide level.

Abstract 5
- NORINE: A PUBLIC RESOURCE FOR NONRIBOSOMAL PEPTIDES -

Caboche Ségolène*[1], Pupin Maude[1], Leclère Valérie[2], Jacques Philippe[2], Kucherov Gregory[1]

- [1]LIFL (UMR USTL/CNRS 8022) - INRIA ~ Villeneuve d'Ascq ~ France - [2]ProBioGEM (UPRES EA 1026), Lille1 University ~ Villeneuve d'Ascq ~ France

Motivation: In micro-organisms, nonribosomal peptide synthesis is an alternative pathway that allows the production of small bioactive peptides from multienzymatic assembly lines called NonRibosomal Peptide Synthetases (NRPSs). The products, called NonRibosomal Peptides (NRPs), show a great diversity in composition, structure and function. They are short (two to about fifty amino acids), but can potentially contain more than 300 different amino acids (instead of twenty amino acids composing regular proteins). The NRP primary structure can be linear like in classical ribosomal peptides, but it is often more complex (totally or partially cyclic, branched or even poly-cyclic). The NRPs harbour a large spectrum of biological activities (e.g. antibiotics, antitumors, immunosuppressors). In spite of a great interest in NRPs due to their particularities and their important bioactivities, few computational resources and dedicated tools are currently available.

Methods: We have developed Norine, a public resource for NRPs. It contains more than 700 peptides and is still growing. Each peptide is annotated with various data collected from scientific publications. Those include the peptide name, its molecular weight, producer organisms, bibliographical references and links to other databases (UniProt and PubChem). The most original information stored in Norine is the NRP structure. We chose to represent the NRP structures at the amino acid level that reflects their biosynthesis, rather than to use the classical chemical representation. Indeed, the NRPSs successively incorporate complete amino acids rather than atoms. A friendly web interface was developed to search for NRPs according to various search criteria. In addition, users can search for a complete structure or a structural pattern (part of a structure possibly with jokers).

Results: Norine is the first resource entirely devoted to NRPs and is available at http://bioinfo.lifl.fr/norine/. We believe that Norine can have various usages in a wide range of related biological studies and can be useful in different applications of NRPs including very important applications in pharmacology. Indeed, we hope that Norine can contribute to biosynthetic engineering efforts to reprogram the NRP assembly lines, in particular because it makes possible systematic studies of the function-structure relationship of NRPs.

Abstract 6
- MOLECULAR DINAMIC OF ACTIVE SITE REGION OF MONILIPHOTHORA PERNICIOSA CHITIN SYNTHASE, THE AGENT OF WITCHES' BROOM DISEASE OF COCOA -

Souza Catiane[1], Taranto Alex*[2], Góes-Neto Aristóteles[3], Sandra Assis[4], Avery Mitchell[5]

- [1]Department of Biological Sciences, 2Graduate Program in Biotechnology (PPGBiotec - UEFS/FIOCRUZ-BA) ~ Feira de Santana ~ Brazil - [2]Graduate Program in Biotechnology (PPGBiotec - UEFS/FIOCRUZ-BA), Graduate Program in Vegetal Genetic Resources (RGV), Department of Health Sciences, State University of Feira de Santana ~ Feira de Santana ~ Brazil - [3]Department of Biological Sciences, 2Graduate Program in Biotechnology (PPGBiotec - UEFS/FIOCRUZ-BA) State University of Feira de Santana ~ Feira de Santana ~ Brazil - [4]1Department of Biological Sciences, 2Graduate Program in Biotechnology (PPGBiotec - UEFS/FIOCRUZ-BA), 3Graduate Program in Vegetal Genetic Resources (RGV), 4Department of Health Sciences, 1-4State University of Feira de Santana, Feira de Santana ~ Feira - [5]5Department of Medicinal Chemistry, 6National Center for Natural Products Research, and 7Department of Chemistry and Biochemistry, 5-7The University of Mississippi, MS, USA. ~ Oxford ~ United States

Motivation: The filamentous fungus Moniliophthora perniciosa (Stahel) Aime & Phillips-Mora is a hemibiotrophic Basidiomycota that causes witches' broom disease of cocoa (Theobroma cacao L.). This disease has resulted in a severe decrease in the Brazilian cocoa production, which changed the position of Brazil in the market from the second largest cocoa exporter to a cocoa importer. Chitin synthases (CHS) converts UDP-N-acetylglycosamine into chitin, the main component of the fungal cell wall. These glycosyltransferases have five different expression levels depending on the fungal life cycle stage. Class III chitin synthases act directly in the formation of the cellular wall and are responsible for most of the chitin synthesis in the cell, and are, therefore, a highly specific molecular target for drugs that could inhibit the growth and development of pathogenic fungi, since CHS is the immediate precursor of chitin and catalyzes an irreversible reaction.

Methods: After obtaining the protein sequence, a model of active site was constructed using Homology Modeling approach. The homologous sequence, with 29% identity, was used as template. The model was constructed by SWISS-MODEL, and refined by a set of Molecular Mechanics (MM) and Molecular Dynamics (MD) calculation, both using ff99 force field and implicit solvent model in Amber 8.0. The quality of resultant model was evaluated by PROCHECK 3.0, ANOLEA, and MD simulations.

Results: Ramachandran plot and MD simulations showed that the model has 98.4% of residues in the most favored regions with thermodynamic stability after 2.0 ns. The complete knowledge about the geometry of active site of CHS can be useful to develop new inhibitors against witches' broom disease

Abstract 7
- GPU ACCELERATED RNA-RNA INTERACTION ALGORITHM -

Rizk Guillaume*[1], Lavenier Dominique[1]

- *[1]IRISA-Symbiose ~ Rennes ~ France*

Motivation: Many bioinformatics studies require the analysis of RNA or DNA structures. Packages like Unafold (Markham, N. R. & Zuker, M. Nucleic Acids Res. 2005; 33, W577-W581) provide many tools to study secondary structures. However, the high computational complexity of these algorithms combined with the rapid increase of genomic data triggers the need of faster methods. Current approaches are (1) designing faster algorithms or (2) parallelizing work on multiprocessor systems. Here, we explore the use of graphics processing unit (GPU) to speed up these kind of computations, which possibly exhibits a higher performance/cost ratio than clusters. It has already been successfully used for the computation of the Smith-Waterman alignment (Svetlin A Manavski, Giorgio Valle BMC Bioinformatics 2008 9-S2). We propose to parallelize on GPU the hybrid function of the Unafold package, which computes the stability of the duplex formed by two RNA sequences.

Methods: For an efficient parallelization, GPU need thousands of independent tasks. Parts taking the most time are found via program profiling and are then re-written in a way to expose parallelism. Our GPU implementation uses both parallelism within a single computation of the algorithm and between several execution of the algorithm across multiple pairs of sequences.

Moreover, to achieve good performance the data needed by the algorithm have to be carefully dispatched in different memory spaces of the GPU, according to their size and their access pattern. Another difficulty comes from the need to reduce to a minimum the if-then-else control instructions of the GPU kernels as the GPU is a SIMD (single instruction multiple data) architecture.

Results: Experiments have been done on an octo-core platform (2*Xeon E5430 2.66GHz, 8 GB RAM) with two NVIDIA Tesla cards. We benchmark our GPU implementation on 26000 pairs of sequences of length 50,50 with one or two cards versus the CPU version of the algorithm from one to eight cores. Total time spent for the complete application are respectively 100, 13.1, 9.8 and 5.3 seconds for 1 core, 8 cores, 1 card and 2 cards. GPU are a competitive alternative : the price of a platform with two Tesla cards is about the same as a platform with 8 processors but with 2.5 times the performance. Similar algorithms are used in a wide array of functions, such as the computation of the secondary structure of a single sequence which might also be parallelizable efficiently.

Abstract 8
- CONTRIBUTIONS OF GC ON GENE EXPRESSION: RECOGNIZING THE ROLES OF GC -

Arhondakis Stilianos*[1]

- [1]*Biomedical Research Foundation of Acedemy of Athens ~ Athens ~ Greece*

Motivation: The effect of GC on expression levels is a topic of considerable evolutionary importance, and also has several practical implications for technologies that quantify expression levels. Several groups have addressed to study the influence of base composition on transcription levels in mammalian genomes observed via genome-wide technologies (sequencing- and hybridization-based techniques). Despite some variability among the reports, especially where they estimate a magnitude for this influence, a persisting trend has emerged: GC-rich genes tend to be expressed at higher levels than GC-poor genes.
Methods: Using publicly available collections of expression data from sequencing- (i.e., EST, MPSS) and hybridization-based (i.e., cDNA and short-oligo arrays) techniques, representing a wide range of human tissues, the contribution of GC on gene expression was investigated. When correlations were estimated using each of the available technologies, they were thoroughly assessed by checking for possible technology-specific or experimental effects. This was achieved by performing simple compositional analyses of the transcriptomes, and by taking into consideration known technology-specific limitations/deficits of each technique, leading to the detection of several cases of unreliable correlations, mostly negative ones.
Results: The cross-platform comparison presented here not only confirmed the persistence of positive correlations between base composition of human genes and expression level, but also detected several technology specific features that affect results. In addition, this work shows that the GC level and the compositional distributions of transcripts represent a very simple tool to assess biases in different technologies; furthermore the essentially invariant GC3 distribution of human genes can be considered as a reliable reference to assess gene representativity, and could play a useful role in biomedical or cross-platform comparison studies. In conclusion, a first conservative lower compositional border of the human transcriptomes is proposed, with mean GC3 of coding transcripts detected within any tissue typically above 55%.

Abstract 9
- CREATION OF A CULTURE COLLECTION DATABASE OF DIAZOTROPHIC AND PLANT GROWTH PROMOTER BACTERIA OF EMBRAPA SOJA -

Higashi Susan*[1], Hungria Mariangela[1], Barcellos Fernando Gomes [1]

- [1]*Soils Biotechnology, Embrapa Soja ~ Londrina ~ Brazil*

Motivation: Culture collection maintenance is a primordial item if there is necessity of using microbial genetic resources. Therefore, these collections operate as ex-situ conservation centers of genetic resources and they are essential for metabolic and genetic diversity exploration. Culture collection can act as service collection keeping microbial resources and offering opportunities as biological material dispatch to research institutions, universities, etc, and information dispatch facilitating the use of microbial resources.

The support to these collections requires information storage of involved microorganisms. Consequently, databases become extremely important to keep the integrity and organization of these information. To use them, they must be well organized in databases with registries and documentation.

In this sense, a project was developed with the aim of creation of a bacterial culture collection of diazotrophic and plant growth promoter bacteria with agribusiness importance. Hence, the purpose of this work is to create a database to organize the information related to these microbial cultures.

Methods:

This database project followed some steps (Conceptual, Logical and Physical Modeling) and specifics techniques. The methodology used proceeded as follows.

At first, the conceptual model was implemented after the study of data related to strains of symbiotic diazotrophic bacteria. The Entity-Relationship approach was selected because it is the most well known conceptual modeling technique.

Second, exhaustive tests of conceptual model was done and then the logical one was constructed.

Third, grounded in the logical model, the physical scheme was implemented with MySQL Database 5.0.16.

Afterwards, the first data were organized to the execution of exhaustive tests of the database. In this sense, the database was approved in all the assays and we could testify that the database structure proposed was really appropriate.

Holding the correct database structure it was possible to organize the database-user interface (web site). The site implementation was done by Renato Camara da Silva from LNCC and it is disposable at www.bmrc.lncc.br.

Results: Analyzing the results we concluded: the database created allowed appropriated information storage and organization, which was essential to supply the necessity of making information available.

Abstract 10
- METAGENOME ANNOTATION: AN OPPORTUNITY FOR UNDERGRADUATE BIOINFORMATICS TEACHING -

Hingamp Pascal*[1], Brochier Céline[2], Talla Emmanuel[1], Gautheret Daniel[3], Thieffry Denis[1], Herrmann Carl[1]

- [1]Biology Department, Mediterranean University ~ Marseille ~ France - [2]Biology Department, Provence University ~ Marseille ~ France - [3]Université Paris Sud ~ Paris ~ France

Motivation: The bottleneck in genomics is shifting from sequencing to annotating, increasing the demand for expert annotators. It is in the interest of research and future job seekers that graduate training anticipates this trend by introducing students to the art of raw sequence annotation.

Methods: We have taken advantage of the increasing amount of metagenomic data publicly available to develop a teaching environment in which undergraduate students are given the opportunity to “turn data into knowledge”. This internet teaching platform, called “Annotathon”, fosters team work and guides apprentice annotators through each step of in-silico analyses, from ORF identification to functional and phylogenetic classification. Generating raw results is an integral part of the exercise, but emphasis is put on their interpretation and critical assessment.

The online format is ideally suited for student involvement outside class whilst allowing instructors to provide annotators with continuous feedback. Communication relies on classical internal forums and chats, but more importantly on an iterative evaluation cycle which allows students to respond to constructive criticism and produce enhanced versions of their annotations.

Results: The 720 students that have taken part in the Annotathon over the past three years have analyzed a total of 23 Mb of ocean microbial DNA, representing 9500 hours of cumulative annotation. The following aspects of the approach appear to significantly contribute to its success:

a) learning by doing: bioinformatics is best introduced by first hand experience; theoretical considerations are easier to grasp once truly familiar with the tools

b) learning by repetition: repeating the analyses on several sequences gives the students the opportunity to experience a wide range of situations, e.g. BLAST report for widespread proteins versus ORFans etc.

c) learning by excitement: according to students, exploring yet unannotated sequences is a major incentive

d) learning from constructive criticism: giving students the opportunity to correct themselves results in noticeable progression over time

The Annotathon environment is available as an open source software, but teams are also welcome to join us on our public server http://biologie.univ-mrs.fr/annotathon/. Ideally as more teams join in, this could lead to an educative distributed annotation jamboree with room for modest scientific contribution.

Abstract 11
- SPRINTS AT GENESILICO - SOFTWARE ENGINEERING TECHNIQUES IN A BIOINFORMATICS LAB -

Rother Kristian*[2], Papaj Grzegorz[2], Feder Marcin[2], Kosinski Jan[2], Koslowski Lukasz[2], Potrzebowski Wojciech[2], Kaminski Andrzej[2], Pawlowski Marcin[2], Kogut Jan[2], Fijalkowski Maciek[2], Gajda Michal[2], Jarzynka Tomasz[2], Tkalinska Ewa[2], Orlowski Jerzy[2], Tkaczuk Karolina[2], Puton Tomasz[3], Musielak Magdalena[3], Koscinski Lukasz[3], Czwojdrak Joanna[3], Milanowska Kaja[3], Kaminska Katarzyna H[3], Osinski Tomasz[3], Domagalski Marcin[3], Kaczynski Jan[2], Figiel Malgorzata[2], Tuszynska Irina[2], Smit Sandra[4], Knight Rob[5], Huttley Gavin A[6], Bujnicki Janusz M[2]

- [2]International Institute of Molecular and Cell Biology ~ Warsaw ~ Poland - [3]Institute of Molecular Biology and Biotechnology, Faculty of Biology, Adam Mickiewicz University ~ Poznan ~ Poland - [4]Centre for Integrative Bioinformatics VU (IBIVU), Vrije Universiteit Amsterdam ~ Amsterdam ~ Netherlands - [5]Department of Chemistry and Biochemistry, University of Colorado ~ Boulder ~ United States - [6]Computational Genomics Laboratory, John Curtin School of Medical Research, The Australian National University ~ Canberra ~ Australia

Motivation: In our everyday research a multitude of programs were written to handle protein sequence analyses. Despite Python as a common programming language, little code was shared among people, and the result was usually a mess. In late 2007, we decided make a concerted effort to create a pipeline for protein family analysis that is documented, tested, and easy to maintain. To implement the project, a series of Sprints – focused two-day programming sessions – was organized.
Methods: In total, 26 people participated: 18 coders, 4 users, 3 technical staff, and 1 correspondent for internal news. Because many undergraduate students attended, the coding was done in pairs of one junior and one experienced programmer. Each Sprint was followed by a two-week cleanup period, where three experienced programmers added Unit Tests and documentation. One Sprint was devoted to bug fixing. Using the Trac ticket system, 101 bug reports were collected, 77 of which could be fixed within the first week. During bug fixing, additional test code was written to make sure the same bugs cannot reoccur. Our software intensively uses PyCogent [1], a Python library that supports many biological applications. During a smaller, three-continent Sprint with the PyCogent developers, the usage of the library was optimized, and cookbook-style documentation for PyCogent could be developed.
Results: This approach benefits from programming in a focused and communicative environment. Experienced programmers write better code because they know it will be read, and students are trained 'on the job'. The code quality benefits from using Unit Tests, ticket systems, and a code repository. As potential users, all participants became familiar with the software long before it was finished, and therefore provided many important suggestions. The outcome is a software pipeline supporting many steps from BLAST/PSI-BLAST queries, creating, updating and filtering alignments, to writing phylogenetic and other reports. The software and source code is available on www.genesilico.pl/python_sprint.

[1] Knight R, Maxwell P, Birmingham A, Carnes J, Caporaso JG, Easton BC, Eaton M, Hamady M, Lindsay H, Liu Z, Lozupone C, McDonald D, Robeson M, Sammut R, Smit S, Wakefield MJ, Widmann J, Wikman S, Wilson S, Ying H, Huttley GA. 2007. PyCogent: a toolkit for making sense from sequence. Genome Biol. 8(8):R171.

Abstract 12
- WHEN DATA INTEGRATION LEADS TO A NEW CONCEPT : THE ORPHAN ENZYMES -

Labedan Bernard[1], Lespinet Olivier*[1]

- *[1]Institut de Génétique et Microbiologie, Université Paris-Sud 11 ~ Orsay ~ France*

Motivation: Despite the current availability of more than two millions of protein sequences, almost 35% of the enzyme activities (EC numbers) defined by the Nomenclature Committee of the International Union of Biochemistry and Molecular Biology are not associated with any amino acid sequence in major public databases.

The presence of so many EC numbers without an associated sequence (orphan enzymes) appears rather surprising at a time where we are inundated by genomic data. Alleviating this problem of orphanity would be very helpful for the difficult task of annotating and/or reannotating genomes. At any rate, there is an urgent need to bridge this unwanted gap between biochemical knowledge and massive identification of coding sequences and we suggest that the whole community could contribute to this task.

Accordingly, we are proposing a dedicated web service to identify the encoding gene for the maximum number of sequence-less enzyme activities.

Methods: We retrieved data from various public databases (UniProtKB, IntEnz, PDB, BRENDA, KEGG) and we have organized them into ORENZA, an efficient relational data warehouse committed to the exploration of the orphan enzyme universe.

Results: To identify the putative sequences associated with orphan enzyme activities, we clearly need the help of a large array of experts. As a result, the ORENZA resource contains a friendly tool allowing people having sound knowledge about specific enzyme activities to make helpful suggestions online. Each suggestion appears as a new item on each EC number's individual files in ORENZA. If several experts agree on the same suggestion, it would be transmitted to the curators of UniProtKB with a high degree of confidence. If experts disagree, their different advices will be published as they have been set.

We hope that ORENZA will help to resolve a few of our startling results: 1, orphan enzymes are present at about the same proportion in every class and subclass of enzyme activities. 2. orphan enzymes are widely distributed in the main functional categories. This is the case, for instance, of a significant number of enzymes that are involved in various metabolic pathways, despite a multitude of groups worldwide that studied them intensively and extensively for many years. 3. Even model organisms contain orphan enzyme activities (e.g. 189 in E. coli, 225 in man).

Abstract 13
- BIOSTATISTICS APPLIED ON FOOD GENOMICS -

Pafundo Simona*[1], Agrimonti Caterina[1], Gullì Mariolina[1], Marmiroli Nelson[1]

- *[1]University of Parma ~ Parma ~ Italy*

Motivation: In recent years, there has been a growing interest in the application of molecular tools to food analysis. This is a call which is shared between consumer and producers, which ask for more safety from one side, and for more transparency on the other. However, the results obtained with different molecular analyses are frequently complex and consequently difficult to interpreter. Therefore, there is also the need of bioinformatics and biostatistics to analyze these data and to extrapolate the significant ones, that could be useful for interpreting on analysis.

Methods: The reported study presents the application of two bioinformatics tools: i) the SPSS® tool was applied on molecular traceability of olive oil and on the determination of a correct procedures for this traceability, like in assessing the influence of the time of storage of olive oil; ii) and the MATLAB® tool was applied on molecular traceability of allergens in food and on the comparison of results obtained with different instruments. In particular, using Amplified Fragments Length Polymorphisms (AFLPs) to the traceability of olive oil, non-parametrical statistical tests can be used to test the significance of results. The analysis was done using the software SPSS 13.0 for Windows v.13.0.1.
The traceability of allergens in food, in particular for almond (Prunus dulcis), by using SYBR®GreenER™ DNA Melting Curve Temperature Analysis was carried out with on an ABI Prism® 7000 Sequence detection system and also with Applied Biosystems 7900 HT Fast Real-Time PCR System. The raw results obtained were directly compared using MATLAB® R2007b.

Results: Biostatistics is helpful and necessary to food genomics, because it can be provide tools to assess the significantly results obtained by using such statistical methods. Results coming from independent analyses and different instruments can also be compared. This fact can therefore increase reliability and accuracy in food analysis.

Abstract 14
- "HARDY–WEINBERG KERNEL": A NEW SIMILARITY MEASURE FOR THE ANALYSIS OF GENETIC DATA IN COMPLEX PHENOTYPES -

Montesanto Alberto[1], Lagani Vincenzo*[2], Di Cianni Fausta[2], Conforti Domenico[3], Passarino Giuseppe[1]

- [1]Department of Cell Biology, University of Calabria ~ Rende ~ Italy - [2]CEntro di Supercalcolo per l'Ingegneria Computazionale (CESIC) - NEC Italia S.r.l. ~ Rende ~ Italy - [3]Dipartimento di Elettronica, Informatica e Sistemistica, University of Calabria ~ Rende ~ Italy

Motivation: Recent technological advances have led to the accumulation of a remarkable bulk of data on genetic polymorphisms. However the development of new statistical and informatics tools for the effective processing of these data has not been equally fast.
Machine Learning literature counts only a few examples of works focused on the development and application of data mining methods specifically devised for genetic polymorphisms analysis, although countless data – mining studies are focused on the analysis of other kinds of genetic data (e.g. gene expression data, proteomic sequences, etc.).
Aim of our work is to define a new similarity measure, the "Hardy–Weinberg kernel", specifically conceived for incorporating prior knowledge during the study of genetic datasets of marker genotypes.
The characteristic of "Hardy–Weinberg kernel" is that the similarity between genetic profiles is weighted by the estimates of gene frequencies at Hardy-Weinberg equilibrium in the population.

Methods: In order to compare the effectiveness of our similarity measure with respect to other "well established" kernels (Linear, Polynomial and Gaussian kernel), we applied Support Vector Machine (SVM) classification algorithms to a real–world dataset (Passarino et al., Hum Hered. 2006; 62, 213-220). The dataset had been collected in order to investigate the influence of the genetic variability of candidate genes on survival at old age. Several classification tasks were defined on the data, according to the analyses reported in the above cited paper. For each classification task SVM parameters were optimized through a cross validation procedure, while relevant features were selected via a forward – stepwise algorithm.

Results: Hardy-Weinberg kernel performances always matched or overcame other kernels performances, when used in conjunction with the forward stepwise feature selection algorithm. Experiments performed without feature selection demonstrated a significant decreasing of Hardy-Weinberg kernel performances. Interestingly, these experiments allowed us to discover the conditions under which our similarity measure is appropriate. In particular, Hardy-Weinberg kernel's poor performances may result from the inclusion of irrelevant genetic polymorphisms with rare alleles. A feature selection method based on such observation is currently under study.

Abstract 15
- CHARACTERIZATION AND ANALYSIS OF THE EXPRESSION PATTERN OF MICRORNAS IN THE GRAPEVINE VITIS VINIFERA -

Piccolo Viviana*[1], Mica Erica[1], Pè Enrico[2], Pesole Graziano[3], Horner David[1]

- [1]Department of Biomolecular Sciences and Biotechnology, University of Milan ~ Milan ~ Italy - [2]Dip. Settore Agraria, Scuola Sup. di Studi Univ. e Perfezionamento S.Anna ~ Pisa ~ Italy - [3]Istituto Tecnologie Biomediche, Consiglio Nazionale delle Ricerche ~ Bari ~ Italy

Motivation: MicroRNAs are small (19-24 nt) noncoding RNAs that play an important role in the regulation of multiple cell events, inhibiting gene expression at the posttranscriptional level by binding target mRNAs that are subsequently degraded or sequestered from translation. Plant microRNA genes are typically transcribed by Pol II to yield polyadenylated primary miRNAs (pri-miRNA). These undergo nuclear cleavage to produce to a stem loop intermediate (pre-miRNA) with specific thermodinamic features. Further processing yields a miRNA:miRNA* duplex with 2 nt 3’ overhangs that enters a cytoplasmic ribonucleoprotein complex which mediates interaction with target mRNAs.

Systematic analyses of micro RNAs and their expression patterns have been performed in only a few plant model species. The availability of the complete genome sequence of the grapevine (Vitis vinifera), has already permitted genome-wide predictions of microRNAs by purely computational methods. Here we present a comprehensive analysis of expression of both mature microRNAs and their primary transcripts in the grapevine using oligonucleotide arrays and next generation sequencing technologies.

Methods: We integrate tanscriptome information derived from high-throughput Illumina SOLEXA and ABI SOLiD sequence tags derived from both polyA+ transcripts and isolated small RNAs with oligonucleotide array data. We are thus able to detect both mature microRNAs and to establish whether genomic loci corresponding to the pre-miRNA are expressed in various tissues.

Results: Using “next generation” sequencing technologies and oligonucleotide arrays, we are able to demonstrate tissue specificity of expression of many microRNA genes and their precursor sequences. In many cases, the unambiguous alignment of sequence tags derived from polyA+ RNA to the genomic sequence allow provisional mapping of primary microRNA transcripts. It is hoped that the approach outlined here will ultimately provide insights into the regulation of processing of primary microRNAs and precursor microRNAs as well as facilitating identification of sequence elements involved in the regulation of transcription of microRNA genes.

Abstract 16
- AUTOMATIC INFERRING DRUG GENE REGULATORY NETWORKS USING COMPUTATIONAL INTELLIGENCES TOOLS -

Floares Alexandru*[1]

- *[1]SAIA - Solutions of Artificial Intelligence Applications ~ Cluj-Napoca ~ Romania*

Motivation: Mathematical modeling gene regulating networks is important for understanding and controlling them, with various drugs and their dosage regimens. The ordinary differential equations approach is probably the most sensible. Unfortunately, this is also the most difficult, tedious, expensive, and time-consuming approach. There is a need for algorithms to automatically infer such models from high-throughput temporal series data. Computational intelligence techniques seem to be better suited to this challenging task than conventional modeling approaches.

Methods: We developed a reverse engineering algorithm - RODES, from Reversing Ordinary Differential Equations Systems (see e.g., Floares, Neural Networks 2008; 21, 379-386) - for drug gene regulating networks. These are gene networks were the regulation is exerted by transcriptions factors and also by drugs. RODES is based on two computational intelligence techniques: genetic programming and neural networks feedback linearization. RODES takes as inputs high-throughput (e.g., microarray) time series data and automatically infer an ordinary differential equations model, discovering the network's structure, and estimating its parameters. The model can be used to identify the molecular mechanisms involved. The algorithm can deal with missing information - some temporal series of the transcription factors, drugs or drug related compounds are missing from the data. For example, an extreme situation is when one wants to model a drug gene regulating and have only microarray temporal series data at his disposal.

Results: RODES algorithm produces systems of ordinary differential equations from experimental or simulated high-throughput time series data, e.g. microarray data. On simulated data, the accuracy and the CPU time were very good - R2 was 0.99 or 1.00 in most experiments, 1 being the maximal R2. In particular, the RODES CPU time is orders of magnitude smaller than the CPU time of other algorithms proposed in the literature.

This is due to reducing the reversing of an ordinary differential equations system to that of individual algebraic equations, and to the possibility of incorporating common a priori knowledge. To our knowledge, this is the first realistic reverse engineering algorithm, based on genetic programming and neural networks, applicable to large drug gene networks.

Abstract 17
- A STRUCTURE-ACTIVITY STUDY OF CEPHALOSPORINS EMPLOYING SUPPORT VECTOR MACHINES -

Antelo-Collado Aurelio[1], Machin-Gonzalez Andy*[1], Hernandez-Diaz Yaikiel[1], Carrasco-Velar Ramon[2]

- [1]Faculty of Bioinformatic, University of Informatic Sciences ~ La Habana ~ Cuba - [2]Center of Pharmaceutical Chemistry ~ La Habana ~ Cuba

Motivation: In 1988, Frere et. al stated that QSAR of ß-lactamic antibiotics were an impossible dream. The cephalosporins pertain to this compounds family and of course, it must be an impossible dream too. In 2003, one of the authors developed a regression and an artificial neural network model of cephalosporins (Carrasco, R., Phd. Thesis, ISBN 978-959-16-0646-4)
Methods: Now, we present a classification model of this compounds type with the same reported sample using Support Vectors Machines. To establish the models, topologic, topographic, quantum chemical and hybrid indices were employed as molecular descriptors of the 100 reported cephalosporins. Both c-svc and ?-svc were evaluated, varying the parameters c, ?, and ?. As kernel, RBF was selected.
Results: The best classification results (100%) were obtained with 11 independent variables and the c-svc machine with different c and ? pair values (1000, 0.1; 10000, 0.5; 1000, 0.5; 100, 0.5; 10, 0.5; 10000, 0.9; 100, 0.9, respectively). The best classification value for six variables was 94% also using c-svc machine. The application is implemented in Java language using the Libsvm library.

Abstract 18
- A NEW TOOL FOR THE PREDICTION OF BIOLOGICAL ACTIVITY USING COMPUTER NETWORK -

Carrasco-Velar Ramon*[1], Antelo-Collado Aurelio[1], Machin-Gonzalez Andy[1], Hernandez-Diaz Yaikiel[1], Prieto-Entenza Julio Omar[1], Rodríguez-León Alexis Rene[1], Pérez-Valdes Yunier Rene[1], Molina-Souto Yania[1], Mejías-César Yuleidys[1], Villaverde-Martínez Julio Antonio[1], Martí-Pérez Ileane[1]

- *[1]Faculty of Bioinformatic, University of Informatic Sciences ~ La Habana ~ Cuba*

Motivation: The University of Informatics Science has been designed to contribute to the informatization of Cuban society. In this sense, the Faculty of Bioinformatics, in cooperation with the Center of Pharmaceutical Chemistry is working together to developing a platform for the prediction of biological activity. The principal address to reach this objective is the optimal utilization of computational resources of the universities and the research centers.

Methods: The proposed system is implemented in Java for multiplatform use, and includes the following modules:

1. Interface adapted from JMOL visualization software.
2. Molecular editor based in JME applet.
3. Database of organic compounds supported on MySQL.
4. Graphic module for structural search in the database.
5. Module for calculation of topologic and topographic descriptors.
6. Module for quantum chemical calculations.
7. Fuzzy Logic module for data mining and to construct models.
8. Support Vector Machines module for data mining and to construct models.
9. Module to reduce the sample size.

Results: All modules are independent and the inclusion in the platform is done by plug-ins. A classification study in a set of cephalosporins using Support Vector Machines and a distributed quantum chemical structure optimization of 20000 compounds, as examples of the possibilities of the platform, are included.

Abstract 19
- ZEBRAFISH INTERACTOME IN ANALYSIS OF DIOXIN TOXICITY -

Alexeyenko Andrey*[1]

- [1]*Stockhiolm Bioinformatics Center , Stockholm University ~ Stockholm ~ Sweden*

Motivation: An integral analysis of environmental effects at the molecular level requires a global view that dynamically reflects functional changes of individual genes/proteins and their interactions. The latter can be observed in a dedicated well-controlled experiment. However, integrating this data into an interactome is hampered by ubiquitous false positive signals.

Methods: We recently created FunCoup – a public database of gene interaction networks. A deeply optimized technology integrated multiple datasets, such as physical protein interactions, mRNA and protein expression, TF and miRNA gene targeting etc. These multiple pieces of weaker evidence fused into confidently predicted interactions of several types (signaling links, co-membership in a protein complex etc.), and are presented on-line as genome-wide networks of eukaryotic organisms, from A. thaliana to H. sapiens (http://FunCoup.sbc.su.se). Thus, FunCoup augmented the notoriously incomplete interactome landscapes and exposed both regulatory and functional sides of gene networks. With a number of tools for graphical and tabular analysis, interaction components can be aligned and studied both inside and across species.

So far, no interactome have been integrated in fish. We employed FunCoup to compute such a network with data from orthologous proteins in eukaryotic organisms.

Another input was a global set of gene expression profiles in the developing zebrafish (days 1, 2, 3, 4, 5 after fertilization). In the 3-way ANOVA experimental design, dioxin-exposed embryos were compared to control samples. Beyond the traditional co-expression analysis, this enabled calculating functional links that reflect network patterns under the toxic versus normal conditions and tracing them during the embryonic development.

Results: Here we present an integral method of augmenting specific datasets with multi-facetted public information collected across many experiments and species. The interactome gained significant confidence and was employed in an analysis of aquatic toxicity in zebrafish. The network analysis scaled from the global view to individual functional components of the affected sub-networks. Due to the detailed time-course observations, one could specifically see original focal areas of the dioxin poisoning and their further propagation. We could also distinguish collective network components and individual genes that proved to be dioxin-resistant.

Abstract 20
- THE MYCOPLASMA CONJUNCTIVAE GENOME SEQUENCING, ANNOTATION AND ANALYSIS -

Calderon-Copete Sandra P.[1], Falquet Laurent*[1], Wigger Georges[2], Wunderlin Christof[2], Schmidheini Tobias[2], Frey Joachim[3]

- [1]Swiss Institute of Bioinformatics ~ Lausanne ~ Switzerland - [2]Microsynth AG ~ Balgach ~ Switzerland - [3]Institute for Veterinary Bacteriology, University of Bern ~ Bern ~ Switzerland

Motivation: The mollicute Mycoplasma conjunctivae is the aetiological agent leading to infectious keratoconjunctivitis (IKC) in domestic sheep and wild caprinae. Although this pathogen is relatively benign for domestic animals treated by antibiotics, it can lead wild animals to blindness and death. This is a major cause of death in the protected species in the Alps (e.g., Capra ibex, Rupicapra rupicapra).

Methods: The genome was sequenced using a combined technique of GS-FLX (454) and Sanger sequencing and annotated by an automatic pipeline that we designed, using several tools interconnected via PERL scripts. The resulting annotations are stored in a MYSQL database. This pipeline is likely to be adaptable to other prokaryotic species.

Results: The annotated sequence is then uploaded into the mollicutes database MolliGen (http://cbi.labri.fr/outils/molligen/) allowing for comparative genomics.
We present the results with several examples of genome comparison and analysis in search for biological targets (e.g., pathogenic proteins).

Abstract 21
- DNA BARCODE ANALYSIS OF FOURTEEN SNAKE CRUDE VENOMS BOUGHT FROM PRIVATE FACILITIES -

CHEN Nian*[1], ZHAO ShuJin[2], HAN LiPing[2], JIAO Yu[3]

- [1]College of Biological Science and Engineering, South China University of Technology ~ Guangzhou ~ China - [2]Department of Pharmacology, General Hospital of Guangzhou Military Command ~ Guangzhou ~ China - [3] Computational Sciences and Engineering Division, Oak Ridge National Laboratory ~ Oak Ridge, TN 37831 ~ United States

Motivation: Correct taxonomic characterization of crude snake venoms has serious implications for the replicability of results from toxicological research. Here, we demonstrate the successful extraction of genomic DNA from dried snake venoms and inferring their probable original species by DNA barcode analysis of the partial mitochondrial 16S gene sequences. Among which one sample has been preserved in a pharmaceutical factory for seven years.

Methods: Extract the genomic DNA from fourteen snake venoms by the method which established in our lab previously. These venoms are labeled from wild Naja atra, Naja kaouthia, Deinagkistrodon acutus and Glodius halys from different proviences in China and Thailand. DNA extracted from the muscles were used as controls. Their 16S gene sequences were amplified and sequenced. All the sequences were submitted to NCBI and compared against their homologous sequences in the GenBank database. The phylogenetic trees were constructed by NJ and MP methods.

Results: Sequence alignment and phylogenetic analysis concluded that these samples indeed contained the correct 16S gene from the respective original snakes. This approach offers an straightforward method for verifying the identity of unknown snake crude venoms.

Abstract 22
- GOMIR: A STAND ALONE APPLICATION FOR HUMAN MICRORNA TARGET ANALYSIS AND GENE ONTOLOGY CLUSTERING -

Zotos Pantelis[1], Papachristoudis George[2], Michalopoulos Ioannis[1], Roubelakis Maria*[1], Pappa Kalliopi[1], Anagnou Nikolaos[1], Kossida Sophia[1]

- [1]Academy of Athens, Biomedical Research Foundation ~ Athens ~ Greece - [2]MIT ~ Cambridge, MA ~ United States

Motivation: MicroRNAs are single-stranded RNA molecules of about 20–23 nucleotides length found in a wide variety of organisms. MicroRNAs regulate gene expression, by interacting with target mRNAs at specific sites in order to induce cleavage of the message or inhibit translation. Predicting or verifying mRNA targets of specific microRNAs is a difficult process of great importance.

Methods: GOmir is a novel stand-alone application consisting of two separate tools: JTarget and TAGGO. JTarget integrates microRNA target prediction and functional analysis by combining the predicted target genes from TargetScan, miRanda, RNAhybrid and PicTar computational tools and also providing a full gene description and functional analysis for each target gene. On the other hand, TAGGO application is designed to automatically group gene ontology annotations, taking advantage of the Gene Ontology (GO), in order to extract the main attributes of sets of proteins.

Results: GOmir represents a new tool incorporating two separate Java applications integrated into one stand-alone Java application. GOmir (by using up to four different databases) introduces, for the first time, miRNA predicted targets accompanied by (a) full gene description, (b) functional analysis and (c) detailed gene ontology clustering. Additionally a reverse search initiated by a potential target can also be conducted. GOmir can freely be downloaded from http://bioacademy.gr/bioinformatics/projects/GOmir .

Abstract 23
- TOWARDS SEMANTIC INTEROPERABILITY OF BIOINFORMATICS TOOLS AND BIOLOGICAL DATABASES -

Pettifer Steve*[1], Sinnott James[1], Thorne Dave[1], McDermott Phil[1], Marsh James[1], Attwood Teresa[2]

- [1]School of Computer Science, Manchester University ~ Manchester ~ United Kingdom - [2]Faculty of Life Sciences, Manchester University ~ Manchester ~ United Kingdom

Motivation: In the biological sciences, the need to analyse vast amounts of information has become commonplace. Such large-scale analyses often involve drawing together data from a variety of different databases, held remotely on the Internet or locally on in-house servers. Supporting these tasks are ad hoc collections of data-manipulation tools, scripting languages and visualisation software, which are often combined in arcane ways to create cumbersome systems that have been customised for a particular purpose, and are consequently not readily adaptable to other uses. For many day-to-day bioinformatics tasks, the sizes of current databases, and the scale of the analyses necessary, now demand increasing levels of automation; nevertheless, the unique experience and intuition of human researchers is still required to interpret the end results in any meaningful biological way. Putting humans in the loop requires tools to support real-time interaction with these vast and complex data-sets. Numerous tools do exist for this purpose, but many do not have optimal interfaces, most are effectively isolated from other tools and databases owing to incompatible data formats, and many have limited real-time performance when applied to realistically large data-sets: much of the user's cognitive capacity is therefore focused on controlling the software and manipulating esoteric file formats rather than on performing the research.

Methods: To confront these issues, harnessing expertise in human computer interaction, high-performance rendering and distributed systems, and guided by bioinformaticians and end-user biologists, we are building re-usable software components that, together, create a toolkit that is both architecturally sound from a computing point of view, and addresses both user and developer requirements. Key to the system's usability is its direct exploitation of semantics, which, crucially, gives individual components knowledge of their own functionality and allows them to interoperate seamlessly, removing many of the existing barriers and bottlenecks from standard bioinformatics analyses.

Results: The toolkit, termed UTOPIA, is freely available from http://utopia.cs.man.ac.uk.

Abstract 24
- CHIPSTER – USER FRIENDLY ANALYSIS SOFTWARE FOR DNA MICROARRAY DATA -

Kallio Aleksi[1], Tuimala Jarno[1], Hupponen Taavi[1], Klemelä Petri[1], Korpelainen Eija*[1]

- [1]CSC -the Finnish IT center for science ~ Espoo ~ Finland

Motivation: DNA microarray data analysis is a fast developing field and most of the new methods are published in the international Bioconductor project (http://www.bioconductor.org/). These methods are freely available, but their use requires knowledge of the R programming language. This is limiting, because microarray researchers typically have a life science background without programming experience. In order to bridge this gap we have created Chipster (http://chipster.csc.fi/), a user friendly analysis software which brings a comprehensive collec¬tion of up-to-date analysis methods within the reach of bioscientists via its graphical user interface.
Methods: Chipster supports Affymetrix, Illumina, Agilent and cDNA arrays and, being a Java program, it runs on Windows, Linux and Mac OS X. The usual analysis features such as preprocessing, statistical tests, clustering, and annotation are complemented with e.g. linear (mixed) models, bootstrapping hierarchical clustering results, and promoter analysis tools. Chipster currently contains almost 100 analysis and visualization tools, and adding new tools is easy. Users can combine and automate frequently used tools into workflows, or use Chipster's ready made wizards. Chipster keeps track of performed analyses and the user can save the analysis history. The analysis scripts can also be viewed at the source code level.

Chipster's graphical client program runs on the user's own computer and the actual analyses are performed on central computing servers. It is also possible to connect external Web Services to the system. The client software utilizes Java Web Start to make installation and version updates as easy as possible. Chipster is available for local installations and it is open source.

As part of the EMBRACE project (http://www.embracegrid.info/), we are currently developing also programmatic access to Chipster. The pipeline offered as a Web service finds differentially expressed genes in Affymetrix data and runs clustering, annotation, and GO and KEGG enrichment analysis for them.
Results: Taken together, Chipster enables more researchers to benefit from the method development in the R/Bioconductor project by offering an intuitive graphical user interface to the analysis tools. The system allows users to save and share workflows, and the installation and version updates are taken care of centrally.

Abstract 25
- 'BRUKIN2D': A 2D VISUALIZATION AND COMPARISON TOOL FOR LC-MS DATA -

Tsagrasoulis Dimosthenis*[1], Zerefos Panagiotis[2], Loudos George[1], Vlahou Antonia[2], Baumann Marc[3], Kossida Sophia[1]

- [1]Bioinformatics & Medical Informatics Team, Biomedical Research Foundation of the Academy of Athens ~ Athens ~ Greece - [2]Biotechnology Division, Proteomics Unit, Biomedical Research Foundation of the Academy of Athens ~ Athens ~ Greece - [3]Protein Chemistry/Proteomics Laboratory and the Neuroscience Research Program Biomedicum Helsinki ~ Helsinki ~ Finland

Motivation: Liquid Chromatography-Mass Spectrometry (LC-MS) is a commonly used technique to resolve complex protein mixtures. Visualization of large data sets produced from LC-MS, namely the chromatogram and the mass spectra that correspond to its compounds is the focus of this work.
Methods: Specifically, the in-house developed 'Brukin2D' software, built in Matlab 7.4, is presented here. It uses the compound data that is exported from Bruker 'DataAnalysis' program, and depicts the mean mass spectra of all the chromatogram compounds from one LC-MS run, in one 2D contour/density plot. Two contour plots from different chromatograph runs can then be viewed in the same window and automatically compared, in order to find their similarities and their differences.
Results: The results of the comparison can be examined through detailed mass quantification tables, while chromatogram compound statistics are also calculated during the procedure.

Abstract 26
- RetroTector online, a rational tool for analysis of retroviral elements in small and medium size vertebrate genomic sequences. -

Sperber Göran[1], Lövgren Anders[2], Eriksson Nils-Einar[2], Blomberg Jonas*[3]

- [1]Physiology Unit, Dept. of Neuroscience, Uppsala University, Uppsala, Sweden - [2]Linnaeus Centre for Bioinformatics, Biomedical Centre, Uppsala University, Uppsala, Sweden - [3]Section of Virology, Dept. of Medical Sciences, Uppsala University, Uppsala, Sweden

Motivation: The rapid accumulation of genomic information in databases necessitates rapid and specific algorithms for extracting biologically meaningful information. More or less complete retroviral sequences constitute 5-50% of vertebrate genomes, also called proviral or endogenous retroviral sequences; ERVs. After infecting the host, these retroviruses have integrated in germ line cells, and have then been carried in progeny genomes for up to several 100 million years. A better understanding of these sequences can have profound biological and medical consequences.

Methods: RetroTector© is a platform-independent JAVA program for identification and characterization of proviral sequences in vertebrate genomes. The full version (Sperber G et al, NAR 2007), requires a local installation with a MySQL database. Although not overly complicated, the installation may take some time. We have now created a “light” version of RetroTector©, (RetroTector online) which does not require specific installation procedures, and which can be accessed via the world wide web.

Results: RetroTector online (http://www.neuro.uu.se/fysiologi/jbgs) was implemented under the Batchelor web interface (A Lövgren et al, unpublished). It allows both file and FASTA cut-and-paste admission of sequences (5 to 1000 kilobases). Jobs are shown in an IP-number specific list. Results are downloadable as text files, and can be viewed with a stand-alone program, RetroTectorViewer.jar (downloadable from the same site), which has the full graphical capabilities of the basic RetroTector© program. Thus, a detailed analysis of any retroviral sequences found in the submitted sequence is graphically presented, and can be exported in standard formats. With the current server, a complete analysis of a 1 Megabase sequence is complete in under 10 minutes. It is possible to mask nonretroviral repetitive sequences in the submitted sequence before analysis, using host genome specific “brooms”. This increases the specificity of the analysis.
Conclusion: RetroTector online is a rational tool for retrovirological and genomic work.

Abstract 27
- A SCIENTIFIC WORKFLOW APPROACH FOR THE INTEGRATION OF BIOINFORMATICS TOOLS -

Han Youngmahn*[1], Cho Yongseong[1], Lee Sang-Joo[1]

- *[1]Korea Institute of Science and Technology Information ~ Taejon ~ Korea South*

Motivation: Thanks to the rapid development of computer science and information technologies, biology work is no longer restricted to test tubes, petri dishes and pipettes. Many questions in biological research may best be answered by using extensive computational tools and resources. In the past decade "Big Science" such as the Human Genome Project has generated a vast knowledge explosion in biological field. The study of bioinformatics which emerged only a century ago has attracted vast attention in biological research by developing and utilizing a huge abundance of computer applications and statistical techniques to acquire, store, organize, analyze and visualize biological data and to facilitate biological research. However, the abundancy of bioinformatics resources brings in common problems, such as heterogeneity and incompatibility. Scientists find it slow, cumbersome and labor-intensive to establish the connections across different resources. The integration of heterogeneous bioinformatics services becomes emergent and of immense importance in this area.

Methods: Many bioinformatics studies usually require sequential analysis. For example, creating a phylogenetic tree using base or amino acid sequences consists of step by step processes, including sequence homology searching using BLAST program, multiple sequence alignment using ClustalW, editing aligned results with biological editing programs such as BioEdit or GeneDoc, and finally, creating phylogenetic trees using tree-building programs such as PHYLIP or PAUP. Thus, bioinformatics workflow system can best be an approach for effective integration of bioinformatics resources and providing seamless interfaces to facilitate bioinformatics analysis works.

Results: We have developed Bioworks system as an automated framework which enables to easily construct and execute the workflow model of complex bioinformatics analysis processes. Bioworks is based on client-server architecture. The client application provides a graphical user interface for constructing a workflow model of complex biological analysis processes and reporting intermediate results of each analysis process. The server engine not just automates the execution of workflow models, but also mitigates any interoperability issues among the bioinformatics services by the predefined data converting rules.

Abstract 28
- GIBA: A CLUSTERING TOOL FOR DETECTING PROTEIN COMPLEXES -

Moschopoulos Charalampos*[1], Pavlopoulos Giorgos[2], Likothanassis Spiridon [3], Kossida Sofia[1]

- [1]Bioinformatics & Medical Informatics Team, Biomedical Research Foundation of the Academy of Athens ~ Athens ~ Greece - [2]European Molecular Biology Laboratory ~ Heidelberg ~ Germany - [3]Department of Computer Engineering & Informatics, University of Patras ~ Patra ~ Greece

Motivation: The study of protein interactions has been vital to the understanding of how proteins function within the cell. In addition, small group of proteins that interact with each other and are stable over time, called protein complexes, are extremely significant units for the harmonic function of the cells and can also provide information about the prediction of unknown proteins that participate in a protein complex.

Recently, new high – throughput methods such as microarrays, yeast two hybrid system, phage display and mass spectrometry generate enormous datasets of protein – protein interactions. Nevertheless, these methods suffer from a large error rate, where many protein interactions that exist in an organism are not recorded and yield many false positives. Moreover, only a small fraction of protein complexes has been experimentally determined due to the disability of these methods to detect all the proteins composing the under question complexes.

For these reasons, the use of computational methods in order to increase the quality of information of the biological methods and to detect protein complexes is essential. Due to the large number of protein interactions, the computational methods use the model of a graph called protein interaction graph. In such a graph, the vertices represent the proteins of an organism and the edges, the interactions between the proteins. Usually, these graphs are undirected and unweighted.

Methods: In this report, we present a new two step methodology for dealing with the protein complex detection problem. Initially, a clustering algorithm is used such as MCL, RNSC or affinity propagation. In the second step, the results are filtered based either on individual or on combination of 4 different methods (density, haircut operation, best neighbour and cutting edge). Our methodology is implemented in a user friendly tool, where the user can choose the algorithm of his preference.

Results: Extensive experiments were performed in 7 different datasets which were either derived from individual experiments or from online databases. Furthermore, we used 5 different methods in order to evaluate, as objectively as possible, the results of our experiments. We compared our method with 4 other algorithms (Mcode, HCS, SideS and RNSC with the filtering proposed from its creators) and we conclude which algorithmic combination produces the best results.

Abstract 29
- TPARVADB: A DATABASE TO SUPPORT THEILERIA PARVA VACCINE DEVELOPMENT -

VIsendi Paul[1], Bulimo Wallace [2], Ng'ang'a Wanjiku [3], Bishop Richard[4], de Villiers Etienne P.*[4]

- [1]1Center for Biotechnology and Bioinformatics, University of Nairobi ~ Nairobi ~ Kenya - [2]US Army Medical Research Unit – Kenya ~ Nairobi ~ Kenya - [3]School of Computing and Informatics, University of Nairobi ~ Nairobi ~ Kenya - [4]International Livestock Research Institute ~ Nairobi ~ Kenya

Motivation: Despite sequencing and annotation of the Theileria parva genome, vaccine and diagnostic development for East Coast Fever have been hindered due to lack of a user-friendly and specific T. parva database. We sought to develop TparvaDB, to provide a comprehensive resource to facilitate research in the development of an ECF vaccine by providing a single user-friendly database of all genome and related data for Theileria parva.

Methods: TparvaDB is based on the Generic Model Organism Database (GMOD) platform. Data was migrated from the original Manatee annotation database, reformatted, and reconfigured to populate TparvaDB. The Apollo annotation workbench and a comparative genomics pipeline were included to add functionality to TparvaDB.

Results: We have developed TparvaDB, an integrated database for T. parva based on GMOD. TparvaDB houses full genome sequences, Expressed Sequence Tags (ESTs), Massivelly Parallel Signature Sequencing (MPSS) data, vaccine candidate gene and other related data. TparvaDB consists of a web page generated using the GMOD web tool, a database implemented in MySQL using the Chado schema. Genomic EST and MPSS data were downloaded from the Manatee annotation database as MySQL dump files and converted into Gene Feature Format (GFF), for loading into Chado. TparvaDB was extended to incorporate the Apollo annotation workbench to facilitate subsequent online annotation. The database was designed to integrate data from other apicomplexan species such as T. annulata and P. falciparum to facilitate for comparative analysis. TparvaDB will greatly enhance the ongoing efforts in ECF vaccine and diagnostic.

Abstract 30
- A GREATER DIVERSITY OF RIBOSWITCHES IDENTIFIED THROUGH THE PRESENCE OF ALTERNATIVE STRUCTURES AND OTHER CONSTRAINTS -

Naville Magali*[1], Marchais Antonin[1], Gautheret Daniel[1]

- [1]Institut de Génétique et Microbiologie, Université Paris-Sud 11 ~ Orsay ~ France

Motivation: Riboswitches are non-coding RNA elements located in 5' untranslated region of genes that control gene expression in response to specific ligands. Currently, efficient methods for riboswitch computational prediction mostly rely on sequence and/or structure conservation. As a likely consequence of this kind of approach, riboswitch families present a marked uniformity in terms of structure, if not sequence, conservation, even between distant species. Here we propose a new and different protocol for the detection of more evolutionary isolated systems, based on their mechanism of action. This approach, which is not limited by conservation criteria, allows the prediction of novel riboswitches that escaped established screening methods.

Methods: Our detection strategy identifies regulatory systems based on a terminator/anti-terminator model. This includes the majority of riboswitches in certain bacterial lignages like Firmicutes, as well as T-boxes or simple attenuators. For now, the prediction was applied to 7 species including Bacillus subtilis, in which the relatively abundant annotation allowed a statistical validation of the method. First, all 5' non-coding regions of genes are extracted and screened for rho-independent terminators. A region encompassing the direct strand of each detected terminator is used as a probe for RNAhybrid, a program that looks for a possible hybridization target in the remaining upstream sequence. The probe/target couple corresponds to a putative anti-terminator structure. Predictions are ranked using a combination of criteria including putative anti-terminator free energy or flanking gene distance, orientation and function.

Results: In Bacillus subtilis, our initial screen led to the detection of 718 5'-terminators, among which 38 correspond to known riboswitch/attenuator systems. The subsequent screens based on free energy and flanking gene information retained 82 candidates among which 32 known riboswitches/attenuators. By increasing significantly the specificity without altering much the sensitivity, our screening procedure thus appear particularly relevant. Many novel candidates are found upstream of transporters or secondary metabolite processing genes. This promising approach should considerably diversify our collection of riboswitches and attenuator systems, notably in rho-independant terminator-rich species.

Abstract 31
- PHD-SNP1.0: A WEB SERVER FOR THE PREDICTION OF HUMAN GENETIC DISEASES ASSOCIATED TO MISSENSE SINGLE NUCLEOTIDE POLYMORPHISMS -

Calabrese Remo*[1], Capriotti Emidio[2], Casadio Rita[1]

- [1]Biocomputing Group, University of Bologna ~ Bologna ~ Italy - [2]Structural Genomics Unit, Department of Bioinformatics, Centro de Investigacion Principe Felipe (CIPF) ~ Valencia ~ Spain

Motivation: Single Nucleotide Polymorphisms (SNPs) are the most frequent type of genetic variation in humans (Collins et al., 1998). Great interest is focused on non-synonymous coding SNPs (nscSNPs) that are responsible of protein single point mutation, since mutations occurring in coding regions may have a larger effect on gene functionality. The possibility of retrieving a large dataset of annotated SNPs from the Swiss-Prot Database (Boeckmann et al., 2003) prompted the application of machine learning techniques to predict the insurgence of human diseases due to single point protein mutation starting from the protein sequence (Capriotti et al 2006).

Methods: We developed a method based on support vector machines (SVMs) that starting from the protein sequence information and evolutionary information, when available, can predict whether a new phenotype derived from a nscSNP can be related to a genetic disease in humans. The system is based on two different SVMs, one is a SVM-sequence that performs predictions relying on sequence information alone, the other is a SVM-profile performing predictions on profile features when evolutionary information is available. Merging in a unique framework the two SVMs we get a hybrid predictive method.

Results: On a recent dataset (April 2008) of 34314 mutations, 48% of which are disease related, out of 7351 proteins, we show that our method can reach more than 72% accuracy (with a correlation coefficient of 45%) in the specific task of predicting whether a single point mutation can be disease related or not. Although based on few informations, our system reaches the same accuracy, with a higher correlation, of the other web-available predictors implementing different approaches (Ramensky et al., 2002 ; Ng and Henikoff, 2003). We design a web server integrating our SVM models, called Predictor of human Deleterious Single Nucleotide Polymorphisms (PhD-SNP). The server is a user friendly resource that gives the possibility of retrieving predictions via e-mail. The submission form is very simple and the user has to paste the query sequence, to select the mutation position and the mutated residue in relative input boxes; furthermore he can choose the predictive method. Best results are obtained when evolutionary information is available and when it is possible to perform predictions using the hybrid predictive method.

Abstract 32
- A CHEMOGENOMICS VIEW OF PROTEIN-LIGAND SPACES -

Helena Strömbergsson*[1], Gerard Kleywegt[1]

- [1]*Uppsala University ~ Uppsala ~ Sweden*

Motivation: Chemogenomics is an emerging inter-disciplinary approach to drug discovery that can be defined as the systematic study of the biological effect of a wide array of small molecular-weight ligands on a wide array of macromolecular targets. The field merges traditional ligand-based approaches with biological information on drug targets and lies at the interface of chemistry, biology and informatics. The ultimate goal in chemogenomics is to understand molecular recognition between all possible ligands and all possible drug targets. However, the size of the protein-ligand space makes any systematic experimental characterization of that space impossible. Protein and ligand space have previously been studied as separate entities, but chemogenomic studies deal with large datasets that cover parts of the protein-ligand space. Since chemogenomics deals not only with ligands but also with the macromolecules the ligands interact with it is of interest to find means to explore, compare and visualize protein-ligand spaces as single entities.

Methods: Two chemogenomic protein-ligand interaction datasets were generated for this study. The first dataset represents the structural protein-ligand space, and includes all non-redundant protein-ligand interactions found in the worldwide Protein Data Bank. The second dataset contains all approved drugs and drug targets stored in the DrugBank database, and represents the approved drug-drug target space. To capture biological and physicochemical features of the chemogenomics datasets, descriptors were computed from the primary sequences of the proteins and the three-dimensional structures of the ligands. Principal component analysis was used to analyze the multidimensional data and to create global models of protein-ligand space.

Results: In this study, we present an approach to visualize protein-ligand spaces from a chemogenomics perspective, where both ligand and protein features are taken into account. The method can be applied to any protein-ligand interaction dataset. Here, the approach is applied to analyze the structural protein-ligand space and the protein-ligand space of all approved drugs. We show that this approach can be used to visualize and compare chemogenomics datasets, and to identify close neighbours in the protein-ligand space.

Abstract 33
- DECIPHERING THE CONNECTIVITY STRUCTURE OF BIOLOGICAL NETWORKS USING MIXNET -

Picard Franck[1], Miele Vincent*[1], Daudin Jean-Jacques[2], Cottret Ludovic[3], Robin Stephane[2]

- [1]CNRS ~ Lyon ~ France - [2]AgroParistech ~ Paris ~ France - [3]Universite Lyon 1 ~ Lyon ~ France

Motivation: Understanding the structure of complex networks has become a challenging task which is tackled using clustering techniques. The principle is to gather the nodes of the network into subsets easier to interpret. Several strategies have been proposed, hubs and modules constituting the two most commonly searched substructures, and for those methods, nodes degree often constitutes the building information. Two major criticisms can be made: hubs and modules constitute only two examples of substructures that can be found in networks, and degree only gives a crude view of the network connectivity.

Methods: We present MixNet, a method dedicated to the analysis of networks connectivity structure. It is based on a clustering procedure that uses mixture models to find groups of nodes sharing similar connectivity patterns without any a priori on the characteristics of the groups. Consequently Mixnet can find modules and hubs, but also other structures such as stars, cliques and product connectivity within the same network, these structures being learned directly from the data. Another advantage of MixNet is that it offers a real opportunity to find structures without defining multicriteria strategies that are not robust.

Results: MixNet is applied to various biological networks. MixNet can be used to summarize and understand the information flow that structures the cortex network. We discuss the finding of hubs in the Cortex macaque network that were previously investigated, show how the method can identify core structure like peripherical and central hubs based on the model only. Then MixNet is used to summarize the regulation diagram of the E. Coli transcriptional network. Interestingly, we show that the connectivity structure revealed by the model reflects the building blocks of the network. We define meta motifs at the group level, such as the feed forward loop which imply global regulators, and discuss the identification of important regulatory nodes from the connectivity point of view. We also show the potential of the method on metabolic and food web networks. Interestingly we find that the summary network which is provided by MixNet reflects the core connectivity structures that build the network, which makes the method a valuable tool to understand the functioning of complex network in a reliable manner. The software package as well as examples are available at http://pbil.univ-lyon1.fr/software/mixnet/.

Abstract 34
- COMPARISON OF ABC TRANSPORTER GENES OF PLASMODIUM SPECIES: A SEARCH IN ELUCIDATING NEW DISCOVERIES TOWARD MALARIA PARASITE ERADICATION -

OLUWAGBEMI OLUGBENGA*[1], Yah Clarence[2], Adebiyi Ezekiel[1]

- [1]DEPARTMENT OF COMPUTER AND INFORMATION SCIENCES, COVENANT UNIVERSITY ~ OTA ~ Nigeria - [2]DEPARTMENT OF BIOLOGICAL SCIENCES, COVENANT UNIVERSITY ~ OTA ~ Nigeria

Motivation: Malaria is a major public health problem associated with high mortality and morbidity rates in Sub- Saharan countries, with a spectrum of systemic complications ranging from mild and self-limiting to life-threatening. Drug resistance has posed a major problem in malaria control and occurs in areas endemic of malaria parasite.

Methods: The current research engaged the use of bioinformatics approach to seek new chemotherapeutic strategies in analyzing and proffering solutions to malaria control and eradication. Three Plasmodium species: P. berghei, P. chabaudi, P. falciparum resistance genes ((ABC transporter) putative genes) were compared using the Atermis comparative tool (ACT).

Results: There was slight variation in the up/down stream alignment within the genes likewise their phylogenetic relationships. This therefore showed that same resistance genes within a population of the same site may vary within the same drug.

Keywords: ABC transporter genes, Plasmodium, eradication

Abstract 35
- THE EFFECT OF SINGLE MUTATIONS ON THE CARRIER ACTIVITY OF THE DICARBOXYLATE CARRIER (DIC) OF S. CEREVISIAE: IN VITRO VALIDATION OF PREDICTIONS OF PROTEIN STABILITY CHANGES -

Ferramosca Alessandra *[1], Mirto Luisa*[2], Tasco Gianluca[3], Tartarini Daniele[2], Zara Vincenzo[1], Aloisio Giovanni*[2], Casadio Rita*[3]

- [1]Di.S.Te.B.A ~ University of Salento, Lecce ~ Italy - [2]SPACI Consortium, University of Salento, Lecce ~ & NNL/CNR-INFM, Lecce ~ Italy - [3]Biocomputing Group ~ University of Bologna ~ Italy

Motivation: A basic problem of structural biochemistry studies is to which extent a mutation will affect the stability, and then the function of the protein. From this point of view, an important aspect regards the function of metabolite mitochondrial carriers, in relation to the role that such proteins cover in some mitochondrial pathologies. Sequence studies have shown that the PX(D/E)XX(K/R) signature is characteristic of all mitochondrial carriers, and possibly involved in the transition from the open to closed states, corresponding to the active/inactive state of the carrier.

In this study our approach is to combine predictions and experimental validation adopting as a test case the dicarboxylate carrier (DIC) of S. cerevisiae. Because DIC structure is unknown, we integrated a routinely expert dependent strategy in a automatic tool based on a Grid infrastructure to facilitate the generation of carrier models, including site directed mutagenesis.

Methods: Transport activity of mutagenized proteins was measured in vitro in reconstituted systems. In parallel in silico experiments protein stability was predicted by using I-Mutant3, available at http://gpcr2.biocomp.unibo.it/cgi/predictors/I-Mutant3.0/I-Mutant3.0.cgi. DIC structure was computed by homology modelling using a service that integrates several software and data involved in a Grid infrastructure, available at https://sara.unile.it/cgi-bin/bioinfo/enter.

Results: Protein stability changes upon mutations can be both predicted and/or tested in vitro, by monitoring the function of mutagenized proteins in reconstituted systems. We found that mutations in the DIC carrier signature reduced or blocked nearly completely the protein transport activity, in agreement with the functional role of the carrier motif. Then the question is whether the observed effect on the activity can be also related to a change on protein stability upon mutation. This was tackled by computational methods. According to our predictions protein destabilization would correlate with the observed loss of protein activity. Our data therefore corroborate the finding that single point mutations may hamper protein stability when placed in functionally relevant structural position, and further add to the possibility of predicting a priori whether the mutation destabilize the protein.

Abstract 36
- IMGT/LIGMOTIF: A TOOL FOR IMMUNOGLOBULIN AND T CELL RECEPTOR GENE IDENTIFICATION AND DESCRIPTION IN LARGE GENOMIC SEQUENCES -

Lane Jérôme*[1], Lefranc Marie-Paule[1], Duroux Patrice[1]

- *[1]IMGT®, the international ImMunoGeneTics information system®, Université Montpellier 2 ~ Montpellier ~ France*

Motivation: The immunoglobulins (IG) and T cell receptors (TR) are the major molecular components of the adaptive immune response of vertebrates. IG and TR loci consist of variable (V), diversity (D) and joining (J) genes organized in multigene groups. Owing to the unusual structure of IG and TR genes, conventional bioinformatic software are not adapted to their identification and description in large genomic sequences. A tool, IMGT/LIGMotif, has been developed for IG and TR gene prediction and annotation. It is based on IMGT-ONTOLOGY, the first ontology in immunogenetics, at IMGT®, the international ImMunoGeneTics information system® (*http://imgt.cines.fr*).

Methods: The software model is based on V, D and J gene prototypes. These prototypes are described by 45 standardized IMGT® labels organized on structural biological criteria. Sixteen labels refer to short specific motifs: splicing sites, heptamers, nonamers and conserved amino acids. They are defined by their conserved sequences and their positions in prototypes. Fifteen longer labels refer to core coding regions, recombination signals and gene units. These long motifs are stored in referential sequence databases. The remaining labels are required for a complete annotation.
The algorithm comprises several steps. (1) The analysed genomic sequence is aligned using BLAST with the referential sequence databases. (2) The obtained alignments (or HSP for High Scoring Pairs) are selected on their E-value, score, identity and length. (3) As the HSPs are defined with labels and as the localization of the labels in genes is defined by prototypes, the next step consists in grouping HSP in genes. (4) Short motifs are searched within and in the neighbouring of labels identified in the previous step. (5) In the last step, short motifs are used as anchors to localized labels precisely and to complete the annotation. At this step IMGT/V-QUEST software is also used for the V genes.

Results: IMGT/LIGMotif was evaluated for gene description and identification in human and mouse IG and TR large genomic sequences (7 megabases). In term of prediction, all known human V, D and J genes were identified by IMGT/LIGMotif. More pseudogenes were found than expected. These sequences are currently analysed to check their biological significance. In term of gene description, the V, D and J gene labels were all correctly assigned.

Abstract 37
- FUNCTIONAL ASSESSMENT OF TIME COURSE MICROARRAY DATA -

Conesa Ana*[1], Nueda Maria José[2], García-García Francisco[1], Sebastian Patricia[1], Dopazo Joaquín[1], Ferrer Alberto[3]

- [1]Bioinformatics and Genomics Department, Centro de Investigación Principe Felipe ~ Valencia ~ Spain - [2]Department of Statistics and Operative Research, University of Alicante ~ Alicante ~ Spain - [3]Department of Statistics and Operative Research, Poliechnical University of Valencia ~ Valencia ~ Spain

Motivation: Time-course microarray study the evolution of gene expression along time across one or several experimental conditions (series). Most developed analysis methods are focussed on the clustering or differential expression analysis of the genes in the dataset and do not integrate functional information. In this work we propose two methods for the functional assessment in time course microarray data that directly exploits the dynamics of expression of the functional categories genes are annotated to.

Methods: We have adapted two methods previously developed for the analysis of time course data, to integrate gene functional information. maSigFun derives from the maSigPro methodology, a regression strategy that models expression patterns and identify genes with patterns differences across experimental series. maSigFun fits a regression model for groups of genes labelled by a functional class and selects those categories which has significant model. ASCA-functional is an extension of the ASCA-genes method. ASCA uses ANOVA and PCA to identify principal components associated to time expression signals. The ASCA-functional strategy uses these leverage values to rank genes which are then analyzed by GSEA (Gene Set Enrichment Analysis). Significant classes are those which are enriched within the genes of best resemblance to the major patterns of time gene expression evolution. We used simulated and experimental datasets. Results were compared with the more traditional approach represented by a linear method followed by enrichment analysis of significant genes. As functional scheme, the Gene Ontology was used.

Results: Simulation studies indicated that classes were positively identified when more than 50% of the annotated genes were correlated. Small size clasess were better detected by maSigFun than ASCA-functional. A cutoff value of 0.4 was optimal for the R2 parameter in the maSigFun. Both maSigFun and ASCA-functional were able to identify more and semantically distinctive functional classes than the comparing method. maSigFun selected specific classes with a good level of internal coherence. Furthermore, the time expression co-expression pattern was of was directly depicted by this methology. ASCA-genes tend to identify classes with a larger number of genes and appeared to be a good adaptation of the Gene Set methods to experiments where more than two conditions are involved

Abstract 38
- ANALYSING GENE EXPRESSION PATTERNS IN THE METABOLIC NETWORK OF NEUROBLASTOMA TUMOURS WITH WAVELET TRANSFORMS -

Schramm Gunnar[1], Gaarz Andrea[1], Seitz Hanna[2], Oswald Marcus[2], Eils Roland[2], Konig Rainer[2]

- [1]IPMB, Bioquant, University of Heidelberg ~ Heidelberg ~ Germany - [2]Institute of Computer Science, University of Heidelberg ~ Heidelberg ~ Germany

Motivation: Neuroblastoma tumours show a very heterogeneous clinical picture ranging from rapid growth with fatal outcome to spontaneous regression or differentiation into benign ganglioneuroma. Specific treatment is crucial and can be supported by understanding the molecular functionality of the tumour.

Methods: We have performed gene expression profiling with microarrays for this tumour and mapped it onto the metabolic network to define biochemical pathways that show a discriminative regulation behaviour between the different tumour types. We used an established method (König et. al., BMC Bioinformatics, 2006) that calculates Haar wavelet transforms on adjacency matrices of the network. These wavelet transforms are normally applied to pattern recognition on images. The method was further developed applying heuristic solutions for the grid arrangement problem.

Results: With this we were able to evaluate all KEGG maps in respect to their ability to discriminate neuroblastoma tumours of patients with favourable and unfavourable outcome. The most significant patterns were found for e.g. purine, pyrimidine metabolism, and one-carbon-pool-by-folate indicating increased nucleotide production for proliferation. Furthermore, we found an interesting significant pattern in the glutamate metabolism indicating a potential switch like behavior of the aggressive tumour. Especially the glutamate and one carbon pool metabolism suit well for further analysis by drug treatment and knock down experiments in the laboratory to define drug targets for the aggressive tumours.

Abstract 39
- INFERRING THE ASSOCIATION OF GENOMIC EXPRESSION AND COPY NUMBER VARIATION -

Orsini Massimiliano[1], Capobianco Enrico[1]

- [1]CRS4 Bioinformatics Laboratory ~ Pula (CA) ~ Italy

Motivation: MicroRNAs are small non-coding RNAs (~22 nucleotides) regulating target gene expression via cleavage or translational inhibition. Lu et al (Nature, 2005) showed that most of microRNAs have differential expression (gain or loss) values in tumour samples, and other studies have mapped tissue-specific cancer signatures (Volinia et al, PNAS, 2006). The location of microRNAs in relation to copy number variation (CNV) has also been recently addressed (Lamy e al, Brit J Cancer, 2006) to reveal possible possible correlation patterns in three types of cancers, but unexpectedly clear signatures could not be established. We hypothesize the same kind of possible associations in brain cancers, and suggest a model representation to specify and test relationships among variables. The rationale is that cancers are also characterized by chromosomal aberration that may be predictive of disease outcome, and many by somatically acquired copy number changes, including loss of heterozygosity (LOH) at multiple loci. These aberration are strongly associated with clinical phenotype including patient outcome.
Methods: The model approach starts from a general genetic signature G which may depend on two general factors, expression X and variation Y, such that $G = A(X,Y) + E$. We thus propose a flexible stochastic model where observable and latent variables can be combined, specified and tested. Nowadays, high throughput array-based methods deliver huge amounts of data for expression, genotyping and CNV leading to a parallel assessment of multiple genomic alterations. We have developed a micro-target warehousing system by tissue, miRWare, aimed to allow coordinated inference, and a tool for automatic annotation of regions highlighted in CNV experiments, Magellano which returns gene structure, SNPs, disease association and expression profiles of each gene in a selected genomic region.
Results: As an example from Magellano, the region 3p14.4-p25.3 (often identified in neuroblastoma) contains 599 genes with 22 microRNAs in part differentially expressed in neural tissue cancer. Brain cancers offer a wealth of data due to the richness of microRNA expression and the tissue-specific stem cell differentiation. We have thus reduced the scale of the analysis compared to Lamy et al who looked at colon, prostate and bladder cancers, and have instead emphasized brain-specific characterization from cell line evidence.

Abstract 40
- EMBOSS ON THE GRID WITH EMBOSS-GUI -

Valverde Jose R*[1]

- *[1]Scientific Computing Service ~ Madrid ~ Spain*

Motivation: The availability of a Grid port of EMBOSS provides a ready solution for some common Bioinformatics tasks requiring large computing power as that provided by the Grid (specially analysis at the genomic level), but at a significant cost in learning the command line. A web interface that hides Grid and EMBOSS complexity can greatly empower users needing to perform these tasks.

Methods: Due to the nature of EMBOSS implementation, the easiest way to make these complex tasks easily usable is to exploit EMBOSS generic command interface,resulting in all EMBOSS applications being adapted for Grid use by acting only at one single level. We have analyzed various solutions to implement a web based GUI for EMBOSS, and finally settled for EMBOSS-Explorer (aka EMBOSS-GUI) as our initial target.

We have analyzed and tested various different implementation approaches, which finally led us to generate a fork of Luke McCarthy original project in order to satisfy Grid policy requirements.

Results: Deciding on an initial target for a web user interface to EMBOSS was a difficult decision were technical and subjective factors. Once we had settled on EMBOSS-GUI as our initial target, the initial adaptation to make it run jobs on the Grid was relatively simple, but in order to accommodate it to Grid policies we had to fork a new project to include user authentication.

Our experience using EMBOSS-GUI for Grid work shows that this solution is wanting in some features required for modern, advanced users and sheds light for future developments using other user interfaces. It also evidences the impact of current Grid policies on the way bioinformatics analysis are normally carried out.

In this work we discuss not only the technical problems faced and the solutions developed but also reflect on how other off line factors such as social, personal and vital issues, usability and policies affect development of modern Bioinformatics solutions.

Abstract 41
- TRANSCRIPTION FACTORS REGULATION OF HUMAN GENE NETWORK -

Krivosheev Ivan*[3], Du Lei[4]

- [3]Department of Bioinformatics, Harbin Institute of Technology ~ Harbin ~ China - [4]Department of Bioinformatics, Harbin Medical University ~ Harbin ~ China

Motivation: With increasing variety of molecular networks, such as regulatory, interaction etc, finding relations between structural features of these networks and their biological significance attracts attention of many researchers. While abundant quantity of studies represented connections between protein networks and topological parameters, few approaches have been applied to transcriptional regulations of coexpressed gene networks. Here, we focus on how the number of transcription factors (TFs) correlate with gene network properties.

Methods: We applied three graph-theoretical characteristics - node degree, betweenness centrality and pairwise disconnectivity - to the analysis of gene coexpression networks from HapMap human gene expression data. The networks were constructed by using ARACNE algorithm. We revealed hundreds of genes, each is subject to massive regulation by several TF. We examine the relationship between topological features of each gene and the number of TFs regulating it with the Spearman coorelation and statistical evaluation.

Results: We demonstrate that for human coexpressed gene networks, betweenness centrality and node degree are negatively correlated with according TF number. Our study provides global insights into the effects of TFs regulation in human gene interactions.

Abstract 42
- GRID-BASED BUSINESS-TO-ACADEMIA COLLABORATIONS -

Kamuzinzi Richard*[1], Bottu Guy[2], Colet Marc[1]

- [1]Bioinformatics Unit, Université Libre de Bruxelles ~ Gosselies ~ Belgium - [2]Belgian EMBnet Node - Bioinformatics Unit, Université Libre de Bruxelles ~ Bruxelles ~ Belgium

Motivation: Modern organizations active in the Life Sciences can no longer envisage research and development (R&D) without the involvement of multiple, distributed and independent partners. Actually, the R&D processes are often complex tasks where resources such as data and analysis services have to be shared and integrated among internal and external collaborative entities. At the same time organizations, especially those from the industrial sector, expect to reach the same or even better research objectives while reducing supporting costs and risks. Thus, in a particular research project, participating partners create a virtual organization (VO) and bring together different scientific disciplines and computing resources to address the underlying goal of the research project. Information confidential to a party must be properly managed to ensure the preservation of intellectual property rights.

Results: In this presentation, we present some of the last outcomes of SIMDAT, a research project funded by the European commission. Within this project, we successfully addressed this need of research virtualisation by developing advanced prototypes to demonstrate the feasibility of online collaborations. Additionally, the prototypes showed that both academic and commercial organizations can build reliable and dynamic partnerships to support collaborative R&D processes. In other words, along with the common business-to-business (B2B) collaboration scheme, organizations in e-science can really also adopt the business-to-academia (B2A) collaboration scheme. The approach adopted to design these prototypes is based on a service oriented GRID infrastructure, namely GRIA, which complies with security constraints imposed by the industrial sector. Moreover, the KDE workflow platform from InforSense is used to access the services infrastructure in order to combine applications and services exposed by different parties. Finally, the web interface wEMBOSS is used to provide the end user (the analyst) with a simple environment from which she/he could interact with the VO without having to worry about the technological level issues.

To conclude, we believe that the EMBnet organization could benefit from SIMDAT experience and developments to build a GRID-based network of industry strength services where different resources could be serviced by different nodes and the whole deployed and viewed as a single organization.

Abstract 43
- THE INTERPRETATION OF PROTEIN STRUCTURES BASED ON GRAPH THEORY -

Habibi Mahnaz*[1], Eslahchi Changiz [1], Sadeghi Mehdy[1], Pezeshk Hamid[1]

- [1]*Faculty of Mathematics, Shahid Beheshti University, Tehran, Iran ~ Tehran ~ Iran*

Motivation: The analysis of protein structure is a challenging problem in bioinformatic allowing detailed exploration of the biological function. There are several features of protein structure which help to predict the protein function.

The main goal of this paper is to understand notions of various geometric aspects of a protein by considering a protein structure as a graph. This approach enables us to calculate important geometric concepts such as packing density and atom accessible surface by investigating the graph properties.

In the current method, for calculating packing density, "Voronoi polyhedron" of a protein is considered. Furthermore, it is possible to define a closed polyhedron on the surface. Various algorithms are used to cause the probe to visit all possible points of contact with the model. The locus of either the centre of the probe or the tangent point to the model is recorded.

Methods: We introduce two new algorithms. The first algorithm creates the maximum polymer inside the sphere of radius R. Using the packing polymers; we determine the packing density of a molecule. The second is based on the graph theory. We determine the hydrophobic cores of a protein as a subgraph which have a large average degree.

Results: For the calculation of the packing density based on the position of a probe sphere, the radius of probe sphere has to approach to zero. But by decreasing the radius of probe ball the time of algorithm rapidly increases. The time complexity of our proposed graph theatrical approach is O(n2), where n is the number of residues of a protein. Applying this algorithm, packing density can be obtained without the calculation of solvent accessible surface and 3D coordinates.

In addition, we present a new algorithm of order O(n2) (where n is the number of atoms of a protein) to calculate molecular surface area of each atom and amino acids. Using these values, we obtained total molecular surface area of a protein and the amino acids which are located in the surface of protein. We show that the packing density value and total accessible surface of a protein are negatively correlated.

Abstract 44
- MOLMETH: THE MOLECULAR METHODS DATABASE -

Lagercrantz Erik[1], Oelrich Johan[2], Martinez Barrio Alvaro[3], Bongcam-Rudloff Erik*[1], Landegren Ulf[2]

- [1]Department of Animal Breeding and Genetics, SLU ~ Uppsala ~ Sweden - [2]Department of Genetics and Pathology, Uppsala University ~ Uppsala ~ Sweden - [3]Linnaeus Centre for Bioinformatics, Uppsala University ~ Uppsala ~ Sweden

Motivation: MolMeth, short for molecular methods, is a database system that catalogs laboratory protocols and methods for the life sciences. It is of particular value for large-scale applications in biobanks and systems biology, but also provides value in scientific communication about molecular procedure in general. It is designed to meet a growing need for structure in protocol specifications while offering convenience for contributors and easy access for end users. Structured protocols offer several advantages over current "flat file" protocol databases, such as allowing protocol presentation be adapted for different purposes. It also provides a foundation for automated reasoning regarding protocols.

Methods: The system presents itself as a web site for searching, retrieving and viewing protocols. Registered users can submit and modify their protocols, and modifications result in new versions with distinct, permanent URL:s. There is also a web service, which allows third party applications to retrieve structured versions of protocols.

The database stores various properties related to each protocol, including a unique accession number, information about materials and available suppliers, versioning information, user comments, references and related entries. The submitter of a protocol is allowed to specify a publication date and rudimentary access rights.

The basic structure of a protocol is modular, meaning that it can be built as a hierarchy of (sub-) protocols, combining steps into different protocols without duplicating common parts. Each protocol is also viewed as a function, which transforms an input to some output, specified using well defined ontologies.

Results: The modularity saves effort for authors when protocols have steps in common, or when a protocol is part of another, more extensive protocol. Protocols that are split into modules are still presented with contiguous instructions in a hierarchical list of steps, adapted for a specific setting if desired.

The MolMeth team hopes that that the computational abilities arising from structured protocols will allow the system to automatically suggest steps for protocol authors or, given a start condition and a goal, even suggest entire protocols by combining smaller protocols from the database. It is already clear that structured protocols will play a role in the development of harmonised standards in several pan-European research infrastructures.

Read more: *www.molmeth.org*

Abstract 45
- A GREEDY ALGORITHM FOR HAPLOTYPE INFERENCE BY PURE PARSIMONY -

Poormohammadi Hadi*[1], Eslahchi Changiz [1], Kargar Mehdi[2], Pirhaji Leila[3], Pezeshk Hamid[4], Sadeghi Mehdi[5]

- [1]Faculty of Mathematics, Shahid Beheshti University, Tehran, Iran ~ Tehran ~ Iran - [2]Department of Computer Engineering, Sharif University of Technology, Tehran, Iran ~ Tehran ~ Iran - [3]Department of Biotechnology, College of Science, University of Tehran, Tehran, Iran ~ Tehran ~ Iran - [4]Center of Excellence in Biomathematics, School of Mathematics, Statistics and Computer Sciences ~ Tehran ~ Iran - [5]National Institute for Genetic Engineering and Biotechnology, Tehran, Iran ~ Tehran ~ Iran

Motivation: Haplotype are important information in the study of complex diseases and drug design. However, due to technological limitations, genotype data rather than haplotype are usually obtained. Thus, haplotype inference from genotype data using computational methods is of interest for many researchers.

Methods: There are several models for inferring haplotypes. One of the most important models is haplotype inference by pure parsimony (HIPP), consisting of finding the minimum number of haplotypes that can resolve all given genotypes. HIPP is an NP-hard problem. In this paper we propose a new greedy algorithm for this problem. The greedy algorithm accurately predicts an efficient Haplotype for inferring the remaining genotypes in each step.

Results: Results of applying our algorithm on a variety of biological and simulated data show that it is very effective with a high accuracy compared to other algorithms.
Also a new measure for evaluating the effectiveness of the algorithms is introduced. This measure is based on the pure parsimony approach which seeks to find the minimum number of haplotypes for resolving the input genotypes.

Abstract 46
- GAINS AND LOSSES OF LINEAGE-SPECIFIC GROUP II INTRON IN MITOCHONDRIA OF GYMNOSPERMS: MOLECULAR EVOLUTIONARY AND PHYLOGENETIC IMPLICATIONS -

Regina Teresa M.R.[2], Quagliariello Carla*[2]

- [2]Dipartimento di Biologia Cellulare, Università degli Studi della Calabria ~ Arcavacata di Rende (CS) ~ Italy

Motivation: The mitochondrial rps3 gene harbours a single group II intron (rps3i1) at a well conserved insertion site from algae up to the angiosperms analyzed so far, with the exception of Beta and Marchantia. Interestingly, in gymnosperms the rps3 reading frame is split by two group II intron, rps3i1 and rps3i2 [Regina et al. J. Mol. Evol. 2005; 60, 196-206]. In this study we surveyed a wide range of representatives of all the extant gymnosperms to get insights into allocation and conservation of group II introns and further test the performance of the novel rps3 intron gains and losses as informative character in phylogenetic inferences among the four living gymnosperm orders (Burleigh et al. Am. J. Bot. 2004; 91, 1599-1613).
Methods: Total genomic DNA was isolated by standard CTAB method (Doyle and Doyle Focus 1990; 12, 13-15) or provided directly by the DNA bank at the Kew Royal Botanic Gardens (UK). The mitochondrial rps3 introns were amplified by PCR using specific primers and directly sequenced. Nuclear (18S), plastidial (rbcL) and mitochondrial (cox1, atpA, rps3) sequences were retrieved from GenBank. Structural alignments of (i.) concatenated mitochondrial sequences and (ii.) concatenated mitochondrial-plastid-nuclear sequences were conducted with ClustalX (Thompson et al. Nucleic Acids Res. 1997; 24, 4876-4882) and used to form a multigene and a multigenome matrix, respectively. Maximum parsimony (MP) and maximum likelihood (ML) analyses were, thus, performed using PAUP* V. 4.0b10 program (Swofford 2003 Sinauer, Sunderland, MA).
Results: We report the shared presence of both rps3i1 and rps3i2 in most of the surveyed gymnosperms but unveil several remarkable exceptions among closely related species. Therefore, we show that the distribution pattern of the rps3 introns is able to discriminate among divergent lineages of living gymnosperms. Furthermore, our multigene and/or multigenome MP and ML analyses demonstrate the mitochondrial rps3i2 as a proper informative character to highlight new mitochondrial genomic endeavours and diverse innovations characterizing the plant molecular biodiversity as well as to reinterpret the phylogenetic inter- and intrafamilial relationships among the extant lineages of gymnosperms.

Abstract 47
- GENE REGULATORY NETWORKS IN BACTERIOPHAGES -

Klucar Lubos*[1], Stano Matej*[1], Hajduk Matus[1]

- *[1]Institute of Molecular Biology SAS ~ Bratislava ~ Slovakia*

Motivation: Complex approach to the study of biological systems is of increasing importance. In order to achieve this task immense amounts of data is needed, which requires computer preprocessing of these data. An important role in this process play biological databases which store records in an easily accessible form for whole scientific community. A system approach to biological data integration and consequent construction of predictive models is the most valuable outcome of systems biology.

Methods: phiSITE database is built upon the MySQL database (version 4.0) and PHP (version 4.3). For visualization of Gene Regulatory Networks BioTapestry program (version 2.1.0, www.biotapestry.org) was used. The simulations of Gene Regulatory networks (GRNs) were run on the Dizzy simulation engine (version 1.11.4, http://magnet.systemsbiology.net/software/Dizzy/).

Results: We have developed phiSITE - database of gene regulation in bacteriophages (www.phisite.org). To date it contains detailed information about almost 500 cis-regulatory elements from 42 bacteriophages. Based on the phiSITE data we defined GRNs for four phages: Enterobacteria phage lambda, Mycoplasma virus P1, Enterobacteria phage Mu and Bacillus phage GA-1 used for visualization in BioTapestry viewer. Next, we created a scaffold of gene regulatory network model of Enterobacteria phage lambda. The model is written in SBML and it is simplified to the level of transcriptional control. We omitted Paq and Pi promoters since they do not influence the simulation significantly. Because phiSITE database does not contain the exact kinetic data of transcriptional processes, these were not specified in the model (experimentally obtained values can be added to the model when desired). We launched the model under two different conditions. In the first test, there were no phage proteins present – values for initial amounts of all protein species were set to 0. In the second test we simulated lysogeny conditions by high initial concentrations of CII (100) and CIII (40) proteins. Both, deterministic and stochastic simulations in Dizzy simulator, produced similar results that correspond with progress of lambda infection in a living bacterial cells even with the lack of exact kinetic data. This fact refers to the robust nature of lambda gene regulation. This work was funded by APVT-51-025044 grant from Slovak Research and Development Agency.

Abstract 48
- THE DIVERGENCE OF EXPRESSION PATTERNS OF DUPLICATED GENES IN ORYZA SATIVA -

Li Zhe[1], Zhang He[1], Gao Ge[1], Luo Jingchu*[1]

- [1]*College of Life Sciences, Peking University ~ Beijing ~ China*

Motivation: Genome-wide duplication is ubiquitous during the diversification of the angiosperms, and gene duplication is one of the most important mechanisms for evolutionary novelties. As an indicator of functional evolution, the divergence of expression patterns following duplication events have drawn great attention in recent studies.
Methods: Here, using large-scale whole-genome microarray data, we systematically analyzed expression divergence of genes arising through whole-genome and small-scale duplication events, in the rice (Oryza sativa ssp. japonica) genome.
Results: Our results shown that duplicates created by whole-genome duplication that retained in colinear segments shown more similar expression patterns than those created by small-scale duplication. We propose that such difference could largely be explained by sequence divergence. Further analysis suggested sequence divergence plays important roles in modeling the divergence of expression patterns, and the mode of duplication had less effect on the divergence of expression patterns.

Abstract 49
- ENGINEDB: A REPOSITORY OF FUNCTIONAL ANALOGUES -

De Sario Giulia*[1], Donvito Giacinto[2], Tulipano Angelica[1], Maggi Giorgio[3], Gisel Andreas[1]

- [1]Istituto di Tecnologie Biomediche, Sede Bari, CNR ~ Bari ~ Italy - [2]INFN Bari ~ Bari ~ Italy - [3]Dipartimento Interateneo di Fisica, Università e Politecnico di Bari ~ Bari ~ Italy

Motivation: Up to now, more than 4,0 million gene products from more than 150000 different species have been described specifying their functions, the processes they are involved in and their cellular localization using a well defined and structured vocabulary, the Gene Ontology (GO). Finding gene products with similar functions or involved in similar biological processes within the same or between different organisms, not relying on the conventional sequence similarity method, is an approach to find analogous gene products, which have similar functions, but not necessarily similar sequences as homologous gene products. However comparing gene products functionalities according to the GO terminology is a very time consuming process.
Methods: ENGINE (gENe analoGue fINdEr) is a tool that parallelizes the search process and distributes the calculation over the computational GRID, splitting the process into many sub-processes (Tulipano et al., BMC Bioinformatics 2007; 8,329-342). We developed a new, more performing version of engine and a process to select the most significant functional analogous gene products. Further, the search results are stored in a relational database (engineDB) hosting the most important information validating the proposed functional analogy between different gene products. A graphical interface enables the user to visualize the proposed functional analogues for his gene product under investigation ordered by the level of calculated analogy. engineDB visualizes the value of the chi-square test we used for the comparison as a rating for the analogy, the GO terms of both compared gene products and the number of GO terms in common and not in common since those are the terms influencing the analogy calculation and important for the user to understand which functionalities made the gene products in comparison more or less similar.
Results: ENGINE has produced for every gene product stored in the GO database a list of potential functionally analogues within and between species using, in place of the sequence, the GO gene description. Those data are publicly available either through a search tool as a GUI to engineDB or as a database dump of the whole data set. The GUI offers to the end user several external links such as UniProt, ENSEMBL and RefSeq and to download specific data and further information such as sequence similarity and protein domain comparison giving a complete overview about the proposed functional analogues.

Abstract 50
- PHYLOGENETIC DATABASE QUERY SYSTEM FOR SPECIES DETERMINATION -

vicario saverio[1]

- [1]*CNR - ITB ~ Bari ~ Italy*

Motivation: Species diagnose is still a complex and knowledge intensive activity. This does not allow society at large to benefit of all advantage of correct systematic information. This limits the recognition of the benefit and interest of systematics. The barcode initiative try to set up a standardized and automatic protocol of species diagnose based on standardized molecular sequences and a database of sequences belonging to known species. Here we implement a protocol that uses explicit phylogenetic inference to treat barcode data for identification. The protocol try to balance the need of fast answer typical of database query tools with the need to have a robust phylogenetic inference that would give answer in term of probability. We explored the efficacy of this protocol under various conditions of speciation and sampling

Methods: The data that we use to test our protocol are the set of the barcode quality sequences available on GenBank/EMBL/DDBJ for lepidopterans. This 3523 sequences were organized in a database grouped few groups based on a priori phylogenetic knowledge. For each group a Bayesian inference based on realistic evolutionary model was performed. A hierarchical query system was build to place an unknown sequence first in the correct group and then in the phylogenetic tree of the chosen group, taking account the uncertainty of the inference. The placement in the pre-computed phylogenetic trees was based on three different methodologies. The system was tested in a cross validation framework and the different topological placement methods compared.

Results: The protocol performed quite well with overall high accuracy (>.95) although error concentrate in species with problematic phylogenetic pattern (polyphyly or paraphyly of the species sequences) or species with very few representatives and distant sister taxa.

in conclusion the methods although rather efficient is very dependent from the phylogenetic pattern for the marker under examination.

Abstract 51
- COMPARING THE BIOCHEMICAL NETWORKS OF HUMAN AND RODENT CELLS INFECTED WITH DIFFERENT PLASMODIUM SPECIES -

Fatumo Segun[1], Schramm Gunnar[2], Adabiyi Ezekiel[1], Eils Roland[2], Konig Rainer[2]

- [1]Department of Computer and Information Sciences, College of Science and Technology, Covenant University ~ Ota ~ Nigeria - [2]IPMB, Bioquant, University of Heidelberg ~ Heidelberg ~ Germany

Motivation: There are about 156 species of Plasmodium which infect vertebrates. Only four of these species infect human: Plasmodium falciparum, Plasmodium vivax, Plasmodium ovale and Plasmodium malariae. Other species infect vertebrates including birds, reptiles and rodents. The four rodent malaria parasites are Plasmodium berghei, Plasmodium yoelii, Plasmodium chabaudi and Plasmodium vinckei. Since there is a high sequence similarity between human and rodents, we have studied the similarities and differences between the parasites that infect these two organisms, in respect to the differences of the hosts.
Methods: In this paper, a computational biochemical approach was employed to identify chokepoints in the four selected species of Plasmodium. A well established method that detects such enzymes in the metabolic networks which uniquely produce or consume a metabolic compound (Yeh. et al., Genome Res., 2004, 14, 917–924) was applied to select these bottlenecks of the networks. These chokepoints were used for discriminating and grouping Plasmodium species.
Results: There existed several common chokepoints enzymes to all the species. We identified an average of 178 chokepoints enzymes in each of these Plasmodium species which are common to all of them. Interestingly, we detected chokepoints which are only common to particular species. These chokepoints helped to partition the parasites into two groups reflecting their dependencies to the hosts. This analysis shows that the differences between the discovered biochemical networks of the Plasmodium species are not only due to lack of knowledge but mainly because of the parasite-host dependencies. Finally, we propose host specific drug targets which have some evidence when compared to the literature.

Abstract 52

- STATISTICAL ASSESSMENT OF DISCRIMINATIVE FEATURES FOR PROTEIN-CODING AND NON CODING CROSS-SPECIES CONSERVED SEQUENCE ELEMENTS -

Creanza Teresa Maria*[1], Horner David S.[2], D'Addabbo Annarita[1], Maglietta Rosalia[1], Mignone Flavio[3], Ancona Nicola[1], Pesole Graziano[4]

- [1]ISSIA-CNR ~ Bari ~ Italy - [2]Dipartimento di Scienze Biomolecolari e Biotecnologie, Universita' di Milano ~ Milano ~ Italy - [3]Dipartimento di Chimica Strutturale e Stereochimica Inorganica, Università di Milano ~ Milano ~ Italy - [4]Dipartimento di Biochimica e Biologia Molecolare, Universita' di Bari ~ Bari ~ Italy

Motivation: The annotation of whole genomes through the identification of coding and regulatory regions is one of the major challenges in the current research in molecular biology. One important topic is identifying the protein coding elements in the set of the mammalian conserved elements. Many features have been proposed for automatically distinguishing coding and non-coding conserved sequence elements (CSEs) making so necessary a systematic statistical assessment of the relevance of single and groups of features in addressing this issue, conditionally to the compared species and to the sequence lengths.

Methods: In our study, we evaluated the relevance of various comparative (based on pairwise cross-genomic comparisons) and intrinsic (based on single-species sequences) features in distinguishing coding from non coding CSEs among human, rat and mouse species by using associative and predictive methods. In order to study the influence of the sequence lengths on the feature performances, the predictive study was performed on different accurately rearranged data sets with coding and non coding alignments in equal number and equally long with an ascending average length. We used Fisher’s linear classifiers trained on single as well as groups of features and estimated their prediction accuracies by using multiple cross validation strategy. The statistical significance and power of the estimated prediction accuracy were evaluated by using non parametric permutation tests. Moreover, by using Kolmogorov-Smirnov non parametric tests we investigated if adding intrinsic features to the comparative ones could improve in a statistically significant way the performances of classifiers.

Results: We found that the most discriminant feature was a comparative measure indicating the proportion of synonymous nucleotide substitutions per synonymous sites. Moreover, linear discriminant classifiers trained by using comparative features in general outperformed classifiers based on intrinsic ones. It results that the combination of comparative features is more powerful in the classification of protein coding sequences while the inverse is true for the intrinsic features independently on sequence length. Finally, the prediction accuracy of classifiers trained by using comparative features increased significantly by adding intrinsic features to the set of input variables.

Abstract 53
- COMPUTATIONAL ANNOTATION OF UTR CIS-REGULATORY MODULES THROUGH FREQUENT PATTERN MINING -

Turi Antonio*[1], Loglisci Corrado[1], Salvemini Eliana[1], Grillo Giorgio[2], Malerba Donato[1] and D'Elia Domenica[2]

- [1]Department of Computer Science, University of Bari ~ Bari ~ Italy - [2]Institute for Biomedical Technologies, CNR ~ Bari ~ Italy

Motivation: The huge amount of data produced by genome sequencing projects has allowed to highlight information on the genetic content of many organisms in the form of lists of genes they can express. Although necessary, this knowledge is not sufficient to understand the mechanisms regulating many events underlying life (i.e., cell growth, differentiation, development). In this sense, it is crucial to decipher the control mechanisms ruling the expression of genome in time and space. To address this problem we have developed a bioinformatic approach based on the use of data mining techniques to detect frequent association of regulatory motifs in untranslated regions (UTRs) of transcripts in Metazoa. The idea is that of mining frequent combinations of translation regulatory motifs, since their significant co-occurrences could reveal functional relationships important for the post-transcriptional control of genome expression.

Methods: The experimentation has been carried out using as a test case UTRs sequences extracted from the MitoRes database, annotated with information available in UTRef and UTRsite databases and collected in a relational database named UTRminer, which supports the pattern mining procedure. The mining approach is two-stepped: first, patterns of regulatory motifs are extracted and annotated in the form of sequences of motifs with information on their sequence location and mutual distances (spacers), then the mutual distances are discretized and the most frequent sequences of motifs and spacers are discovered by means of an algorithm for sequence pattern mining. Frequent sequences have a support greater than a user-specified threshold and the procedure for the generation of frequent sequences is guaranteed to be complete.

Results: The UTR sequences analysed concern ten different species. The total number of analysed sequences is 3896, among which 1944 5’UTRs and 1952 3’UTRs. Frequent motifs patterns, generated at first step, have a complexity ranging from 2 to 3 (number of distinct motifs detected on the same UTR) in 5’UTRs and from 2 to 5 in 3’UTRs. Preliminary results based on the observations and comparative analysis of discovered sequential pattern add new insights to our knowledge about post-transcriptional regulatory mechanisms controlling genome expression, while demonstrating the effectiveness of the bioinformatics approach presented in supporting discovery of motifs patterns.

Abstract 55
- IMGT/V-QUEST: AN ALGORITHM FOR IMMUNOGLOBULIN AND T CELL RECEPTOR SEQUENCE ANALYSIS -

Brochet Xavier*[1], Lefranc Marie-Paule[2], Giudicelli Véronique[1]

- [1]IMGT, LIGM, IGH, UPR1142 ~ Montpellier ~ France - [2]IMGT, LIGM, IGH, UPR1142 ~ Montpellier ~ France

Motivation: The molecular synthesis of the immunoglobulin (IG) and T cell receptor (TR) is particularly complex and unique since it generates an extraordinary diversity of the IG and TR repertoires (1012 antibodies and 1012 TR per individual) which results from several mechanisms at the DNA level: the combinatorial diversity of the variable (V), diversity (D) and joining (J) genes, the N-diversity and, for IG, the somatique hypermutations. IMGT/V-QUEST has been developed for the standardized analysis of IG and TR nucleotide sequences.

Methods: IMGT/V-QUEST identifies the closest V, D, J genes and alleles using pairwise alignment and comparison to expertly annotated and standardized data from the IMGT reference directory which is based on IMGT-ONTOLOGY. The algorithm proceeds in 3 steps for the V genes and alleles identification. (1) it identifies a model sequence by aligning the user sequence to a set of the IMGT reference directory comprising ungapped germline V gene sequences (gaps according to the IMGT numbering are stored for the next step), without allowed insertions or deletions. (2) it gaps the user sequence with the positions of the stored IMGT gaps of the model sequence. (3) it identifies the closest germline genes and alleles by the highest similarity score between the gapped user sequence and the complete IMGT reference directory. An optional step detects potential insertions and deletions in the user sequence by Smith and Waterman alignment with the closest germline genes and alleles. If insertions/deletions are detected, the steps for V gene identification are performed again. The J genes and alleles identification proceeds in 2 steps: the beginning of the J is determined by alignment with the IMGT reference directory. Then the closest germline J genes and alleles are identified by similarity evaluation. At last, the algorithm integrates IMGT/JunctionAnalysis for a detailed analysis of the V-J and V-D-J junctions and an accurate D genes and alleles identification.

Results: IMGT/V-QUEST provides a standardized, complete and accurate characterization of the rearranged IG and TR nucleotide sequences. IMGT/V-QUEST is widely used for the study of the IG and TR repertoires and for antibody engineering. It has been recommended by the European Research Initiative on chronic lymphocytic leukemia (ERIC) for the evaluation of the V genes mutational status.

Abstract 56
- HTC FOR ASPIC: A DISTRIBUTED WEB RESOURCE FOR ALTERNATIVE SPLICING PREDICTION AND TRANSCRIPT ISOFORM CHARACTERIZATION -

D'Antonio Mattia[1], Paoletti Daniele[1], Carrabino Danilo[1], D'Onorio De Meo Paolo[1], Sanna Nico[1], Castrignano' Tiziana[1], Anselmo Anna[2], D'Erchia Anna[3], Licciulli Flavio[4], Mangiulli Marina[3], Mignone Flavio[2], Pavesi Giulio[2], Picardi Ernesto[3], Riva Alberto[5], Rizzi Raffaella[6], Bonizzoni Paola[6], Pesole Graziano[3]

- [1]CASPUR ~ Rome ~ Italy - [2]University of Milan, Dipartimento di Scienze Biomolecolari e Biotecnologie ~ Milan ~ Italy - [3]University of Bari, Dipartimento di Biochimica e Biologia Molecolare ~ Bari ~ Italy - [4]Istituto Tecnologie Biomediche del Consiglio Nazionale delle Ricerche ~ Bari ~ Italy - [5]Department of Molecular Genetics and Microbiology, University of Florida ~ Gainesville ~ United States - [6]DISCo, University of Milan Bicocca ~ Milan ~ Italy

Motivation: Alternative splicing (AS) affects the great majority of intron-containing genes and thus is a major mechanism in the expansion of transcript and protein complexity in eukaryotes. Recent descriptions of the functional implications of AS in tissue-specificity, different biological processes and tumor development has generated an explosion of interest and activity in this field.

Methods: In order to analyse the transcriptome and proteome complexity of multicellular organisms and detect the genes specifically involved in human health and disease, we developed a software platform for high-throughput large-scale alternative splicing analysis and transcript isoform characterization.

This platform, HTC for ASPic, provides independent, flexible and scalable high-throughput large-scale alternative splicing analysis and transcript isoform characterization. It integrates computational intensive algorithms we developed previously [Castrignanò et al. Nucleic Acids Res. 2006 Jul 1;34(Web Server issue):W440-3.][Castrignanò et al. Bioinformatics. 2008 Apr 3.] with suitable web services and databases.

Results: The software system has been optimized programming multi-threaded powerful Java client for data preprocessing and several distributed application servers for intensive computation. HTC for ASPic divides the input into parallel tasks without dependency and therefore it scales linearly with the number of processors. The system is also fault-tolerant.

The web resource is available free of charge for academic and non-profit institutions.

Abstract 57
- GIBBS FREE ENERGY CHANGES OF BIOCHEMICAL REACTIONS INFERRED FROM REACTION SIMILARITIES -

Rother Kristian*[1], Hofmann Sabrina[2], Bulik Sascha[2], Hoppe Andreas[2], Holzhuetter Herrmann-Georg[2]

- [1]Laboratory of Bioinformatics and Protein Engineering, International Institute of Molecular and Cell Biology ~ Warsaw ~ Poland - [2]Computational Biophysics Group, Institute of Biochemistry, Charite Universitätsmedizin ~ Berlin ~ Germany

Motivation: An indispensable prerequisite for the thermodynamic and kinetic modeling of biochemical reaction networks is to assign a reliable value for the standard Gibbs free energy change (DeltaG0) to each reaction and transporter. However, for genome-wide metabolic networks experimental DeltaG0 values are scarce. Here we propose a novel computational method to infer the unknown DeltaG0 value of a reaction from known DeltaG0 values of chemically similar reactions.

Methods: To quantify the chemical similarity of biochemical reactions we have established a detailed classification procedure that assigns 3304 different chemical attributes to atomic groups occurring in presently characterized biochemical metabolites. Changes in these attributes between the substrate and product molecules are tracked on a per-atom basis and similarities between these reaction-specific attribute changes are assessed by the Tanimoto coefficient (T) assuming values between 0 (complete dissimilarity of reactions compared) and 1 (identity of reactions compared).

Results: Testing our method across a set of 1546 biochemical reactions 216 of which being covered by experimentally determined DeltaG0 values - the root-mean-square distance (RMSD) between predicted and measured DeltaG0 values amounted to 8.0 kJ/mol, if a minimum similarity of T>0.6 to reactions with known DeltaG0 values is assumed. This value is significantly smaller than the RMSD of 10.5 kJ/mol achieved with the commonly used group contribution method. However, for less similar reactions, the group contribution method produces a more accurate predictions and a combination of both approaches is proposed. Clustering all reactions of a given metabolic network according to chemical similarity allows to identify minimal sets of reactions for which DeltaG0 values yet have to be experimentally determined in order to make reliable predictions of DeltaG0 values for the remaining reactions.

Abstract 58
- IS IT POSSIBLE TO PREDICT PROTEIN BEHAVIOR IN HYDROPHOBIC INTERACTION CHROMATOGRAPHY AND AQUEOUS TWO-PHASE SYSTEMS USING ONLY THEIR AMINO ACID COMPOSITION? -

Salgado J. Cristian*[1], Asenjo Juan A.[1], Andrews Barbara A.[1]

- [1]Centre for Biochemical Engineering and Biotechnology, Department of Chemical Engineering and Biotechnology, University of Chile ~ Santiago ~ Chile

Motivation: The prediction of the partition behavior of proteins in hydrophobic interaction chromatography (HIC) and aqueous two-phase systems (ATPS) using mathematical models based on amino acid composition was investigated. Predictive models were based on the average surface hydrophobicity (ASH), which is estimated by means of models that require the 3D structure of proteins and by models that use only the amino acid composition. These models were evaluated in a set of 12 proteins with known experimental retention time in HIC and 11 with known partition coefficient in ATPS. Our results indicate that the prediction based on the amino acid composition is feasible for both separation systems, even though the quality of the prediction depends strongly on the operational conditions. In the case of ATPS the best results were obtained by the model which assumes that all of the amino acids are completely exposed. An increase in the predictive capacity of at least 54% with respect to the models which use the 3D structure of the protein was obtained in this case. However, best prediction in HIC was obtained by the model based on a linear estimation of the amino acidic surface composition. This model required additional tuning, but its performance was 5% better than that obtained by the 3D structure model.

ACKNOWLEDGEMENTS: FONDECYT PostDoctoral Research Project 3070031.

KEYWORDS: Amino acid composition; Mathematical model, hydrophobic interaction chromatography and aqueous two-phase systems.

Abstract 59
- IMPROVING THE PREDICTION OF PROTEIN BEHAVIOR IN HYDROPHOBIC INTERACTION CHROMATOGRAPHY AND AQUEOUS TWO-PHASE SYSTEMS WITH CLUSTERING METHODS -

Ugarte Jorge E.[1], Andrews Barbara A.[1], Salgado J. Cristian*[1]

- [1]Centre for Biochemical Engineering and Biotechnology, Department of Chemical Engineering and Biotechnology, University of Chile ~ Santiago ~ Chile

Motivation: The aim of this study is the improvement of mathematical models used to predict the behavior of proteins in hydrophobic interaction chromatography (HIC) and aqueous two-phase systems (ATPS) based on their amino acid composition. This problem was tackled by carrying out clustering analysis over a large database of amino acid properties (APV): Self Organizing Maps, k-means, Simulated Annealing, Growing Neuronal Gas, Growing Grid, and hierarchical Clustering were used. This analysis allows us to generate new APVs from those found in literature, which were used to improve prediction models. Three of these models require only the amino acid composition of proteins and different assumptions regarding the tendency of the amino acids to be exposed to the solvent; the other requires the three dimensional structure of the proteins. These models were adjusted using the new APVs and were evaluated in a set of 12 proteins with known experimental retention time in HIC and 11 with known partition coefficient in ATPS. We found that the best APVs were generated by the Growing Neuronal Gas algorithm. In fact, two vectors that significantly improve the performance of the prediction models were found. Using these vectors the prediction performance of the model based on the 3D structure and the best model based on amino acid composition were improved by 38% and 31%, respectively.

ACKNOWLEDGEMENTS: FONDECYT PostDoctoral Research Project 3070031 and Millennium Scientific Initiative ICM P05-001F.

KEYWORDS: Clustering, prediction models, hydrophobic interaction chromatography and aqueous two-phase systems.

Abstract 60
- MODELING HETEROCYST PATTERN FORMATION IN CYANOBACTERIA -

Gerdtzen Ziomara P.[1], Salgado J. Cristian*[1], Osses Axel[3], Asenjo Juan A.[1], Rapaport Ivan[3], Andrews Barbara A.[1]

- [1]Millennium Institute for Cell Dynamics and Biotechnology, Department of Chemical Engineering and Biotechnology, University of Chile ~ Santiago ~ Chile - [3]Millennium Institute for Cell Dynamics and Biotechnology, Department of Mathematical Engineering, Center for Mathematical Modeling (UMI 2807-CNRS), University of Chile ~ Santiago ~ Chile

Motivation: In this paper we study the process by which vegetative cells of cyanobacteria differentiate into heterocysts in the absence of nitrogen. We propose a simple network which captures the complexity of the differentiation process and the role of all variables involved in this cellular process. Specific characteristics and details of the system's behavior such as transcript profiles for ntcA, hetR and patS between consecutive heterocysts are studied. The proposed model is able to capture one of the most distinctive features of this system: a characteristic distance between two heterocysts, with a small standard deviation according to experimental variability. The system's response to knock out and over expression of patS and hetR was simulated in order to validate the proposed model against experimental observations. In all cases, simulations show good agreement with reported experimental results. The model also shows that refractability of heterocysts to the action of PatS is not required in order to achieve the characteristic differentiation pattern observed in cyanobacteria.

ACKNOWLEDGEMENTS: Millennium Scientific Initiative ICM P05-001F.

KEYWORDS: cyanobacteria, heterocyst, mathematical modeling, cell differentiation and gene network.

Abstract 61
- IDER ASSOCIATED PHYSIOLOGICAL NETWORK IN MYCOBACTERIA -

Ranjan Sarita[1], Ranjan Akash*[2]

- [1]1LEPRA-Blue Peter Research Centre, ~ Hyderabad ~ India - [2]Sun Centre of Excellence in Medical Bioinformatics, EMBnet India Node, CDFD, ~ Hyderabad ~ India

Motivation: Transcription regulators play an important role in coordinating gene expression of physiologically related genes in an organism. Each transcription factor recognizes specific DNA sequence close to promoter regions and modulated gene expression by binding to these sequences. Using bioinformatics, we can learn the DNA sequence pattern associated with these transcription factors and predict genome wide targets of these regulators. Taking operonic context of the predicted targets it is possible construct a model of physiological interaction network that reveals the genome encoded biology associated with the transcriptional regulator. We have taken this approach to model IdeR (iron dependent regulator) associated physiological network in mycobacteria.

Methods: Using a profile based approach the DNA sequence pattern associated with IdeR regulator was recognized. The IdeR associated DNA binding pattern was subsequently used to predict genome wide targets of IdeR in sequenced genomes of mycobacteria. Taking operonic context of the predicted targets we construct a model of physiological interaction network of the IdeR. These interactions were modeled using Cytoscape.

Results: A model IdeR (iron dependent regulator) associated physiological network in mycobacteria constructed. Using comparative genomics approach we have also explored the relative conservation IdeR associated physiological network in different sequenced species of mycobacteria.

Abstract 62
- CORYNEDB: A DATABASE OF IN SILICO IDENTIFIED OPERONS AND TRANSCRIPTIONAL UNITS OF CORYNEBACTERIA -

Ranjan Sarita*[1], Savala Narendra Kumar [2], Ranjan Akash[2]

- [1]LEPRA-Blue Peter Research Centre, ~ Hyderabad ~ India - [2]Sun Centre of Excellence in Medical Bioinformatics, EMBnet India Node, CDFD ~ Hyderabad ~ India

Motivation: Corynebacteria are a diverse group of microorganism belonging to the phylum Actinobacteria. Species belonging to Corynebacterium genus are gram positive, non motile acid fast staining bacilli. Corynebacteria are found in a wide range of different ecological niches such as vegetables, soil, cheese smear, skin, and sewage. Some of the species, such as Corynebacterium diphtheriae and Corynebacterium jeikeium, are important pathogens while others, such as Corynebacterium glutamicum, are of immense industrial importance as they are extensively used to produce amino acid using fermentation process. Corynebacteria are also closely related to other important group of pathogens called mycobacteria.

Methods: We have applied our in house developed operon prediction approach (Ranjan et al BMC Bioinformatics 2006; 7(Suppl 5): S9) to identify operons and transcriptional unit in sequenced genomes of corynebacteria. This approach involved orientation analyses, Intergenic distance analyses, transcriptional terminators analyses and conserved gene cluster analyses.

Results: We have predicted transcriptional units and operons in six sequenced genomes of corynebacteria. The predicted operons and transcriptional units are organized as relational database called CoryneDB. CoryneDB has information about corynebacterial genes and in silico predicted transcriptional units and operons. This database would assist the scientific community, to hypothesize functional linkages between operonic genes of corynebacteria, their experimental characterization and validation.

Abstract 63
- DOOPSEARCH: A WEB-BASED TOOL FOR FINDING AND ANALYZING COMMON CONSERVED MOTIFS IN THE PROMOTER REGIONS OF DIFFERENT CHORDATE AND PLANT GENES -

Sebestyén Endre[1], Nagy Tibor[2], Barta Endre*[2]

- [1]Agricultural Research Institute of the Hungarian Academy of Sciences ~ Martonvásár ~ Hungary - [2]Bioinformatics Group, Agricultural Biotechnology Center ~ Gödöllo ~ Hungary

Motivation: The comparative genomic analysis of a large number of orthologous promoter regions from chordate and plant genes revealed thousands of conserved motifs. Most of these motifs are different from any known transcription factor binding sites (TFBS). To identify new potential TFBSs with in silico analysis methods, we need a tool to be able to search among the conserved motifs. The result of a given search is expected to provide a list of genes, which are associated with a certain conserved motif and might be regulated by a transcription factor recognising the motif and binding to it. To test and confirm the association of a conserved motif with certain types of genes, we can perform a Gene Ontology (GO) analysis on the gene list.

Methods: Based on our DoOP database, we used different taxonomic groups to extract conserved motifs either from the human genome annotation based chordate or the Arabidopsis thaliana based plant database. We have developed a C program called MOFEXT, for performing gapless alignments and fast searches in the different motif collections. The FUZZNUC program from the EMBOSS package has also been implemented to search in the promoter sequences of the DoOP database. We slightly modified the GeneMerge program to use it for the GO analysis of the results. To handle the web-based queries efficiently, all data are stored in a MySQL database. We have developed several PERL modules (available from CPAN) to carry out the querying of the MySQL database, the MOFEXT searches, the GO analysis and the graphical presentation of the results.

Results: We have developed a new web page called DoOPSearch (http://doopsearch.abc.hu) for the analysis of the conserved motifs in the promoter regions of chordate or plant genes. We used the orthologous promoters of the DoOP database to extract conserved motifs from different taxonomic groups. The advantage of this approach is that different sets of conserved motifs may be found depending on how broad the taxonomic coverage of the underlying orthologous promoter sequences is (consider e.g. primates vs. mammals). The DoOPSearch web page allows the user to search these motif collections or the promoter regions of DoOP with user supplied query sequences or any of the conserved motifs from the DoOP database. The gene lists obtained can be further analyzed using the modified GeneMerge program.

Abstract 64
- BRAGOMAP - A NEW PERL SCRIPT FOR HIGH THROUGHOUTPUT BLAST RESULTS ANALYSIS INCLUDING GO AND MAPMAN AUTOMATIC ANNOTATIONS -

Woycicki Rafal*[1], Gutman Wojciech[2], Przybecki Zbigniew[1]

- [1]Department of Plant Genetics, Breeding and Biotechnology, Faculty of Horticulture and Landscape Architecture, Warsaw University of Life Sciences ~ Warsaw ~ Poland - [2]Department of Biochemistry, Faculty of Agronomy and Biology, Warsaw University of Life Sciences ~ Warsaw ~ Poland

Motivation: Analyzing of sequences similarities is the first and most important method used to find out the function of unknown nucleotides.

Searching of homologs should be done carefully not to loose any important ones.

Having thousands of results from various long-read sequencing projects (genomic polymorphons or BAC ends), the by-hand ability to retrieve interesting (to our goal) similarities in hundreds of Blast results decreases rapidly.

Decreasing the number of retrieved sequences by giving more stringency in e-value threshold or displaying less results could lead to false deductions.

Functional genomics, proteomics and metabolomics could give us answers to the role of nucleotide sequences. It makes the need to annotate as much of the homologies as we can, to proper molecular function, biological process and cellular component (as its proposed by widely accepted Gene Ontology Consortium annotations or MapMan mappings by Max-Planc-Institute).

Methods: To facilitate fast retrieval of interesting Blast homologies and making right deductions about the biological role of sequences, in big sequencing projects, the new Perl script BRAGOMAP was written. The program make use of some of BioPerl modules as well as the power of regex text-mining in the Perl itself.

Results: The script gives us the possibility to find interesting sequence similarities by using keywords and giving points for each one found. It collects all important information from the GenBank data and puts it in different columns of tab-delimited file for further use.

If we were interested (for example) in flower differentiation genes we could use the keywords (flower, ovule, anther, etc.) and/or filter all the homologies isolated from flower tissues in a special development stage. We can also filter results by choosing similarities to interesting genes or protein products. This script retrieve also all standard information from the Blast and GenBank files as Description, ACC no., E-value, Similarity positions, Query Length, Percent of Similarity etc.

Automatic GO and MapMan annotations are done by looking for genes, protein products and /or DB references in the proper mappings files.

Here we present the usefulness of the script in analyzing sequence similarities and annotations mapping of 3855 BAC ends obtained from the HindIII BAC genomic library of cucumber (Cucumis sativus L., line B10).

Abstract 65
- IMPROVING GENE RANKING METHODS USING FUZZY CLUSTERING AND GENE ONTOLOGY -

Mohammadi Azadeh*[1], Saraee Mohammad Hossein [1]

- [1]*Advanced Database Syetems, Data Mining and Bioinformatics Research Laboratory, Department of Electrical and Computer Engineering,Isfahan University of Technology ~ Isfahan ~ Iran*

Motivation: Microarray technology allows simultaneously monitoring the expression levels of thousands of genes. An important analysis task for this data is identification of the genes which are significant or mostly associates with a disease. Identification and selection of such genes from thousands of genes in microarray experiments, is called gene selection. A variety of approaches have been proposed for gene selection. A group of methods are ranking methods which measure the discriminatory power of each gene according to a test-statistic and select the genes with highest score as discriminatory genes. Although these kinds of methods have low computational complexity they have two drawbacks. First, they don't consider the correlation between genes and consequently the selected genes have redundancy. Second these methods don't utilize the biological knowledge.

Methods: In this paper we have hybridized fuzzy clustering and gene ontology with ranking methods such as t-test fisher, information gain and TNOM, to solve the mentioned problems. In the proposed method genes are clustered based on their gene expression profiles and their biological knowledge obtained from gene ontology, after that a test statistic is applied on genes to rank them. Since the genes in a cluster represent similar genes, we allow only a limited number of genes from a cluster to be selected, in this way the correlation amongst selected genes decrease considerably.

Results: We have applied the proposed method on colon dataset and compared the result of our method to some ranking methods such as t-test, fisher, information gain and TNOM. The results show that Coupling clustering and gene ranking methods can identify gene subset that has lower redundancy and improves classification accuracy, compared with simple gene ranking methods.

Abstract 66
- GENEFINDER: "IN SILICO" POSITIONAL CLONING OF TRAIT GENES -

Martínez Barrio Álvaro*[1], Lagercrantz Erik[2], Burvall Sofia[3], Bongcam-Rudloff Erik[2]

- [1]The Linnaeus Centre for Bioinformatics, Uppsala University, Biomedical centre, P.O. Box 598, SE-75124 ~ Uppsala ~ Sweden - [2]Department of Animal Breeding and Genetics, Swedish University of Agricultural Sciences, Biomedical centre, P.O. Box 597, SE-751 24 ~ Uppsala ~ Sweden - [3]Uppsala University ~ Uppsala ~ Sweden

Motivation: Positional cloning of trait genes is very laborious and the amount of information on gene function in different organisms is increasing so rapidly that it is hard for a research group to collect all relevant information from a number of data sources without performing a big number of tedious, non-automatized and time consuming searches.

Methods: A specific application known as GeneFinder has been designed and implemented in order to collect information of a trait locus controlling a specific phenotype mapped to a certain chromosomal region. The information retrieved consists of information on gene function, disease conditions, tissue expression and even predicted gene homologies on several other species. All this amount of information is then further combined with a specially devised ranking algorithm. Finally, the results of the search are presented to the user in a friendly ranked list doing easy its interpretation. To achieve this, a web interface to the GeneFinder webservice was also developed. Our distributed application is publicly available and free to use.

Results: We show how our implementation works, both with its web interface and programmatically with the webservice API, in a very short number of steps when searching an example candidate region. An online server for the web interface is available at http://genefinder.ebioinformatics.org/.

Abstract 67 - CONTACT COORDINATION PATTERNS AND ELECTROSTATIC POTENTIAL AT ALPHA CARBON ATOMS: A DOSSIER OF PROTEIN SECONDARY STRUCTURE ELEMENTS -

Borro Luiz[1], Mazoni Ivan[1], Alvarenga Daniel[1], Cecilio Pablo[1], Grassi Jose[1], Jardine Gilberto[1], Mancini Adauto[1], Neshich Goran*[1]

- [1]*Embrapa Informatica Agropecuaria ~ Campinas, SP ~ Brazil*

Motivation: The process of protein folding might be investigated by analyzing the secondary structure elements (SSE) in light of the physical chemical characteristics of amino acids. Molecular structure prediction is also dependent fundamentally on how much we know about the SSE and their interplay within the 3D constellation. Our major motivating factor to study in detail the relationship of selected physical chemical parameters and the capacity of determined sequences to build specific SSE came from the fact that we succeeded to build the most extensive database of such parameters - the STING_DB. This allows us to make "signal enhancement" of patterns hidden within diversity of sequences capable of generating the very same SSE.

Methods: We present here analysis of pre-calculated values for the electrostatic potential at the alpha carbons and for the cross-links, previously stored in the STING_RDB. Our procedure was to first separate the proteins from the PDB according to their classes: all alpha, all beta, alpha + beta and alpha/beta. All SSE were grouped and aligned with respect to their length. All aligned SSE, were then analyzed in terms of 47 sequence/structure descriptors (grouped in 32 major classes) such as: electrostatic potential, sequence conservation, hydrophobicity, accessibility, dihedral angles, internal contacts etc.

The Cross Links are defined as contacts (any type from possible 5 classes: Hydrophobic, Hydrogen Bonding, Aromatic Stacking, Salt bridging and Cystein-bridging) established among residues that are far apart in the protein primary sequence, but are close in its 3D fold. The order of cross link is identified as a number of such cross-links established among independent stretches of sequence (the size of which was fixed to 30 Amino Acids). The higher the order, the more important that residue must be for the protein folding/stability.

Results: We found a clear tendency for EP at alpha carbon atoms for alpha helices and beta strands, having negative and positive values, respectively. We also found a clear and opposing tendency for the value of cross links for alpha helices and beta strands, showing less and more, respectively, cross links in comparison to the other parts of proteins.

Abstract 68
- INTEGRATING ERV SEQUENCE AND STRUCTURAL FEATURES WITH DAS AND EBIOX -

Martínez Barrio Álvaro*[1], Lagercrantz Erik[2], Sperber Göran O[3], Blomberg Jonas[4], Bongcam-Rudloff Erik[2]

- [1]The Linnaeus Centre for Bioinformatics, Uppsala University, Biomedical centre, P.O. Box 598, SE-75124 ~ Uppsala ~ Sweden - [2]Department of Animal Breeding and Genetics, Swedish University of Agricultural Sciences, Biomedical centre, P.O. Box 597, SE-751 24 ~ Uppsala ~ Sweden - [3]Department of Neuroscience, Physiology, Uppsala University ~ Uppsala ~ Sweden - [4]Section of Virology, Department of Medical Sciences, Uppsala University ~ Uppsala ~ Sweden

Motivation: The Distributed Annotation System (DAS) is a protocol used to exchange biological information. The network distribution concept of the protocol makes possible the use of different DAS reference and annotation servers to combine biological sequence data with annotations in order to depict an integrated view of the data to the final user.

Methods: Here we present a DAS annotation server devised to provide information about the endogenous retroviruses (ERV) detected and annotated "in silico" by a specialized tool called RetroTectorTM. We describe the procedure to implement the necessary DAS 1.5 protocol commands to construct DAS annotation servers. We use our server to exemplify those steps. The data distribution is separated from visualization which is carried out by eBioX, a general and user-friendly open-source programme suite with multiple bioinformatics utilities.

Results: We apply the server to discuss the advantages of distributing ERV data using the DAS protocol. Some well characterised ERVs are shown for two different organisms. By doing this, we also demonstrate the modularity of a distributed protocol like DAS as a solution for combining annotations belonging to different species. Reference and annotation data servers are then used in combination with eBioX to provide a friendly visualization of ERVs as well.

Abstract 69
- THE RHNUMTS COMPILATION -

Attimonelli Marcella*[1], Lascaro Daniela[1], Castellana Stefano[1], Gasparre Giuseppe[3], Romeo Giovanni[3], Saccone Cecilia[1]

- [1]Department of Biochemistry and Molecular Biology "E.Quagliariello" ~ Bari ~ Italy - [3]Unit of clinical genetics ~ Bologna ~ Italy

Motivation: Introduction
To a greater or lesser extent, eukaryotic nuclear genomes contain fragments of their mitochondrial genome counterpart, deriving from the random insertion of damaged mtDNA fragments. NumtS (Nuclear mt Sequences) are not equally abundant in all species, and are redundant and polymorphic in terms of number of copies. In population and clinical genetics, it is important to have a complete overview of NumtS quantity and location. Searching PubMed yields hundreds of papers regarding Human NumtS compilation. A comparison of published compilations clearly shows significant discrepancies among data, due both to unwise application of Bioinformatics methods and a not yet correctly assembled nuclear genome. To optimize quantification and localization of NumtS, we have produced a consensus compilation of Human NumtS obtained by applying various bioinformatics approaches.

Methods: Location and quantification of NumtS have been achieved by applying database similarity searching methods: different methods, Blastn, MegaBlast and BLAT, changing both parameters and database have been used. To verify the in silico predicted NumtS we are amplifying and sequencing them from DNA samples of different mitochondrial haplogroups.

Results: The obtained results have been compared, further analysed and checked against the already published compilations thus producing the Reference Human Numt Sequences (RHNumtS) compilation. The resulting NumtS are 190. At present we have sequenced 40 NumtS, those with lower score, thus demonstrating the efficiency of our in silico protocol.
Conclusions
The RHNumtS compilation represents a highly reliable reference basis on which designing a lab protocol to test the reality of each NumtS. We are designing the RHNumtS Compilation database structure for implementation in the HmtDB resource (www.hmtd.uniba.it) .

Abstract 70
- STORING, CLASSIFICATION AND BETA-LACTAMASE SEQUENCE MOLECULAR PATRONS DETECTION -

Rodríguez Luis[1], Ramón Mantilla[1], Falquet Laurent[2], Reguero Maria Teresa[1], Emiliano Barreto*[1]

- *[1]Bioinformatics Center, Biotechnology Institute, National University of Colombia ~ Bogotá ~ Colombia*

- *[2]Swiss Institute of Bioinformatics, Genopode, UNIL, CH-1015 ~Lausanne ~Switzerland*

Motivation: Recently, beta-lactamases have gained attention due to their role in bacteria resistance to beta-lactamic antibiotics. Beta-lactamase identification and classification are big challenges for the molecular epidemiology research, due to the few differences in their DNA sequences. Consequently, is difficult to classify new sequences in the beta-lactamases classes and families already described. The currently used classification is based in the Lahey Clinic and the Pasteur Institute bioinformatics network. Despite this, a bioinformatics system focused on storing outbreaks, sequences, phylogenetic and bibliography data isn't available. The current work was intended to develop it.
Methods: The bioinformatic system BLA.id was implemented in Linux Suse 10 box, using MySQL and including Perl and PHP written scripts and some C ++ programs. The user web interface was built using PHP, MySQL and R. The system automatically retrieved information from international databases such as UniProt, EMBL and PubMed. BLA.id data was manually annotated and reviewed. The classification used was based on lists from Lahey Clinic (http://lahey.org/studies/), Pasteur Institute (http://www.pasteur.fr/recherche/genopole/PF8/betalact_en.html) and single international journal reports.
Results: We were able to develop a system that has beta-lactamases sequences actually reported properly classified. It allows the direct report of new sequences into the data base and their accurate classification. Our bioinformatics system has leaded us to the detection of some classification inconsistencies in beta-lactameses' sequences already reported. It also has permitted the phylogenetic classification and the identification of conserved patterns that are useful for molecular epidemiologic approaches. Moreover, some problems such as the existence of sequence synonyms can be easily solved on this structured solution. BLA.id is accessible via web on http://co.embnet.org/BLA.id/.

Abstract 71
- DETECTING HYDROPHOBIC CLUSTERS IN 3D STRUCTURES AT ATOMIC LEVEL -

Alexeevski Andrei[1], Sergei Spirin*[1], Karyagina Anna[2]

- [1]Belozersky Institute, Moscow State University ~ Moscow ~ Russia - [2]Gamaleya Institute of Epidemiology and Microbiology ~ Moscow ~ Russia

Motivation: The hydrophobic effect is one of the main natural forces stabilizing structures of macromolecules and their complexes. The basis of the hydrophobic effect is that nonpolar atoms tend to dispose together in a water environment.
Methods: Our approach is based on the supposition that every set of closely located nonpolar atoms leads to a force trying to keep their close location. This suggestion dictates an "atomic" level of investigating hydrophobic effect, because, for example, such polar aminoacid residues as lysine, arginine, or glutamate contain nonpolar atoms, which can participate in hydrophobic interactions with other residues or ligands.
Results: We introduce an algorithm for detecting clusters of nonpolar atoms in structures of macromolecules, and a web service implementing this algorithm. The web service is available at http://monkey.belozersky.msu.ru/npidb/cgi-bin/hftri.pl. The program can be used in two modes. First, it can detect hydrophobic clusters in an entire structure, for example, the main hydrophobic core of a protein globule and minor hydrophobic areas at the surface of a molecule. Second, as its input it can take two parts of a macromolecular complex (e.g., two interacting protein molecules) and detect hydrophobic clusters at the interface of those parts. The latter can help to investigate hydrophobic interaction between two molecules.

Also we suggest a procedure to detect conserved hydrophobic cores in a family of related proteins and conserved hydrophobic interactions in a family of related complexes. Programmatic implementation of the procedure is a subject of our current efforts.

We used our program to describe conserved hydrophobic clusters in a number of structural families (e.g., homeodomains and V-set domains), for analysis of inter-subunit interaction in capsids of icosahedral viruses, for analysis of DNA-protein interaction in a number of complexes of DNA-binding proteins with DNA (a part of results was published in: Karyagina et al. J. Bioinf. Comp. Biol. 2006; 4(2); 357-372).

Abstract 72
- TOWARDS BARCODE MARKERS IN FUNGI: AN INTRON MAP OF ASCOMYCOTA MITOCHONDRIA -

Santamaria Monica *[1], Vicario Saverio [1], Domenica D’Elia [1], Pappadà Graziano [2], Scioscia Gaetano [3], Vicario Saverio [1], Scazzocchio Claudio [4], Saccone Cecilia [5]

- [1] CNR - Istituto di Tecnologie Biomediche, Sede di Bari ~ Bari ~ Italy - [2] Exhicon srl ~ Bari ~ Italy - IBM Italy S.p.A. - [3] IBM Innovation Lab ~ Bari ~ - [4] Institut de Gènètique et Microbiologie, UMR 8621 CNRS, Universitè Paris-Sud (XI) ~ Orsay cedex ~ France - [5] Dipartimento di Biochimica e Biologia Molecolare, Università di Bari ~ Bari ~ Italy

Motivation: A rapid, standardized and cost-effective identification system is now essential for Fungi owing to their wide involvement in human health and life quality. Currently the molecular identification of species in Fungi is based primarily on nuclear DNA, but the potential use of mitochondrial markers has also been considered, due to their peculiar and favourable features, among which, above all, their high copy number, the possibility of an easier and cheaper recovering and the paucity of repetitive regions. Unfortunately, a serious difficulty in the PCR and bioinformatic surveys is due to the presence of mobile introns in almost all the fungal mitochondrial genes. The aim of the present work is to verify the incidence of this phenomenon in Ascomycota and to identify one or more mitochondrial gene regions where introns are missing so as to propose them as species markers (barcodes).

Methods: The general trend towards a large occurrence of introns in the mitochondrial genome of Fungi has been confirmed in Ascomycota, except for some specific regions, by an extensive bioinformatic analysis, performed on 7234 records of 11 mitochondrial protein coding genes and 2 mitochondrial rRNA coding genes belonging to this phylum, available in Genbank. A new query approach, developed within a databases federation system designed to manage, integrate and enhance connections among the information possibly hosted in heterogeneous data sources, has been applied to retrieve, in an effective manner, relevant information usually present in the entries of a biological database, but hardly selectable through the classical query systems. This approach has allowed to avoid a series of alignment and retrieval stages based on the similarity calculation, which inevitably produce false positives and negatives in the final results.

Results: Despite the large pervasiveness of introns in Ascomycota mitochondrial genes, the results of the present work have shown that specific regions from at least three alternative genes, namely ND2, ND4 and ND6, seem intron-poor and large enough to be considered barcode candidates for Ascomycota. This finding could be the first step towards a mitochondrial barcoding strategy similar to the standard approach routinely employed in metazoa, and its use would prevent other efforts to look for alternative, less efficient or more expensive, strategies to bypass the introns problem.

Abstract 73 *- RegExpBlasting, a regular expression rules based algorithm to classify a new sequence. -*

Rubino Francesco[1], Attimonelli Marcella[1]

- [1]*Department of Biochemistry and Molecular Biology, Bari, Italy*

Motivation: One of the most frequent usages of bioinformatics tools concerns the functional characterization of a newly produced nucleotide sequence (a query sequence) by applying Blast or FASTA against a set of sequences (the subject sequences). However, in some specific context it could be useful to compare the query sequence against a cluster such as a multialignment. The purpose of the RegExpBlast tool is i) to associate to each multialignment a pattern, defined through the application of regular expression rules; ii) to automatically characterize the submitted nucleotide sequence on the basis of the function of the sequences described by the pattern better matching the query sequence.

Methods: Regular expressions are tools used to represent every possible character variation in a sequence alignment, much more powerful than the classic consensus sequences that leaves out a great quantity of information about that site, like possible SNPs, insertions and deletions.
The RegExpblasting tool is organized in two sections: A and B. RegExpBlasting section A produces the RegExp pattern describing each of the considered multialignments by extracting the regular expression which represents all the variations of the group of sequences available through their multialignment. RegExpblasting section B allows the classification of a new sequence by comparing it to each of the patterns defined in section A and reports as output the matching multialignments where the new sequence can be easily added.
The algorithm section A scans all the multialignment sites and produces a regular expression according to the below described criteria:
1) if a site contains only one type of nucleotide, only this will be included in the regular expression;
2) if the types of nucleotides in a site are two or more, all this nucleotides will be enclosed in a character class to represent every nucleotide variation;
3) if a site contains one or more gaps the meta character “?” will be added to represent this case.
The resulting pattern represents every sequence composing the multialignment.
Section B compares the query sequence with the patterns defined in section A and produces as output the multialignments associated to the best matching patterns.

Results: An application of these algorithms is used in the "characterize your sequence" tool available in the PPNEMA resource [1] http://www.ppnema.uniba.it. PPNEMA is a resource of Ribosomal Cistron sequences from various species grouped according to nematode genera. PPNEMA allows the retrieval of plant nematode multialigned sequences or the classification of a new nematode rDNA sequence by applying RegExpblasting. The same algorithm supports automatic updating of the PPNEMA database also.
References

F.Rubino, A.Voukelatou, F.De Luca, C.De Giorgi and M.Attimonelli “PPNEMA: a database of the RNA cystron from Plant Parasitic nematodes.” Int.J.Plant Genomics, Sp.Issue on Bioinforamtics, 2008, in press.

Abstract 74
- Possible Role for Proximity of Genes in Their Expression in Rice and Arabidopsis -

Shahmuradov Ilham A.[1,2], Akbarova Yagut Yu.[3], Aliyev Jalal A.[2], Qamar Raheel[1], Chohan Shahid Nadeem[1,4], Solovyev Victor V.[5]

- [1]Dept. of Biosciences, COMSATS Institute of Information Technology, Islamabad, Pakistan - [2]Bioinformatics Laboratory, Institute of Botany, Baku, Azerbaijan - [3]College of Medicine and Health Sciences, Sultan Qaboos, University, Muscat, Sultanate of Oman - [4]School of Natural Sciences, University of Western Sydney, Australia - [5]Dept. of Computer Science, Royal Holloway, University of London, Egham, UK

Motivation: In contrast to prokaryotes, the proximity of genes in eukaryotic genomes has not previously been known to play any significant role in their expression profiles. However, two recently reported phenomena in eukaryotes including human, mouse, yeast and a few plant species indicate such a role. These phenomena are: (1) transcription of the Head-to-Head (H2H) adjacent genes from the shared promoter; and (2) chimeric mRNAs and proteins produced via alternative transcription termination, splicing and translation of the Tail-to-Head (T2H) neighboring genes. To further verify this evidence at the genomic scale, we searched through the genomes of rice and Arabidopsis for the presence of H2H and T2H gene pairs at a distance of less than 800 bp and 1000 bp, respectively.

Methods: Gene ontology data for rice and Arabidopsis were obtained from the genome annotations and TAIR WEB-site, (ftp://ftp.Arabidopsis.org/home/tair/Genes/Gene_OntologyATH_GO.20031202.txt), respectively. Analysis of the genome annotations of rice and Arabidopsis was performed using computer programs ARGAN and OSGANn specially developed by us for this task. The pair-wise comparison of amino acid sequences has been carried out by BLAST program. Search for promoters and statistically significant open reading frames were performed by TSSP and BESTORF programs, respectively (*www.softberry.com*).

Results: In the rice genome, 580 H2H and 1,386 T2H gene pairs were found, while 1,898 H2H and 6,618 T2H gene pairs were found in the Arabidopsis genome. The short spacers between H2H genes in both the genomes may serve as potential bidirectional promoters, though they did not reveal any significant evolutionary conservation, However we found striking conservation of intergenic region between the same gene pair in Arabidopsis and Brassica napus with the experimentally confirmed bidirectional promoter. Further studies suggest that “non-stopping” transcription and alternative splicing of some of these T2H pairs may result in chimer transcripts. We obtained cDNA support for 106 and 105 rice and Arabidopsis T2H pairs, respectively, to be transcribed into chimeric or read-through mRNA(s). Analysis of the protein coding potential of such putative transcription-induced chimer genes revealed putative chimer proteins having significant similarity with known plant proteins.

Abstract 75
- A bioinformatics platform to store and analyze alternative splicing events detected by Exon Arrays. -

Licciulli Flavio[1], Picardi Ernesto[2], Pesole Graziano[1,2], Calogero Raffaele[3], Delle Foglie Gianfranco[1], Grillo Giorgio[1], Liuni Sabino[1]

- [1]Istituto Tecnologie Biomediche, Consiglio Nazionale delle Ricerche, Bari, Italy - [2]Dipartimento di Biochimica e Biologia Molecolare "E.Quagliariello", Università di Bari, Bari, Italy - [3]Bioinformatics and Genomics unit, Dipartimento di Scienze Cliniche e Biologiche, Università di Torino, Orbassano, Italy

Motivation: Current studies suggest that more than 90% of human multi-exon genes are subjected to alternative splicing, a key molecular mechanism in which multiple transcripts may be generated from a single gene. Such transcripts may encode proteins with different biological functions or may be subjected to different post-transcriptional regulations through variants of 5' and 3'UTRs. As a consequence, the alternative splicing greatly increases the complexity of the human transcriptome and proteome.

Last generation microarray platforms such as exon arrays now offer a more detailed view of the gene expression profile providing information on the alternative splicing pattern.

Exon arrays, in fact, have a capacity of more than six million data points and have been designed to interrogate approximately one million exons including all annotated exons (from RefSeqs) in addition to computationally and empirically identified exons (including those supported by ESTs and gene predictions as well).

Exon arrays significantly differ from traditional microarray platforms in the number and placement of the oligonucleotide probes, and allow researchers to perform two different but related analyses at both gene and exon level. In particular, the accession to genome-wide exon data can provide answers to challenging issues concerning the identification of tissue-specific exons, biomarkers and isoforms involved in different pathological mechanisms.

In this context and in order to investigate tissue-specific, developmental stage-specific as well as disease-related alternative splicing events, we posed our attention to the GeneChip Human Exon 1.0 ST Array system by Affymetrix. One significant feature of this innovative Affymetrix platform is that individual probes are designed along the full-length of human genes. Moreover, an exon-level probe set is made of 4 probes and more than one exon-level probe set may fall in the same exon. All generated probe sets related to a specific gene are then grouped in transcript clusters. Annotation details such as genome coordinates for each feature and the relationships among exons, probe sets and transcript clusters are provided and frequently updated by Affymetrix.

Current protocols to manage and analyze Affymetrix Exon Array data are not completely defined. Affymetrix proposed a workflow based on pre-filtering of the expression data, transformation of exon-level intensity data in gene-level normalized values called splice index (SI) and statistical validation based on an ANOVA based method based on measuring differences between exon level signal and aggregate gene level signal called MiDAS (Microarray Detection of Alternative Splicing).. Unfortunately, exon array workflows do not include facilities to provide a simple and basic visualization of the putative alternative splicing events (ASEs) to facilitate the interpretation of their biological significance.

With the aim to fill this gap and, thus, provide a user-friendly representation of alternative splicing events in different experimental conditions we built a robust bioinformatics platform based on a data-warehouse approach.

Methods: According to data-warehouse standards suggested by Kimball and Inmon (Inmon 1995. Prism Solutions), our bioinformatics platform can manage a huge quantity of data and, in the meantime, facilitate their biological interpretation by using Business Intelligent (BI) techniques. In its first version, the platform integrates experimental data produced by Affymetrix Exon 1.0 ST Arrays and related annotations from several public and specialised databases such as AspicDB (Castrignanò et al. Nucleic Acids Res. 2006; 34,W440-443), Gene, GO and KEGG. At the heart of our system is the mapping of Exon Array probe sets along all transcripts predicted and annotated in ASPicDB in order to provide powerful and stimulating insights at genomic and transcriptomic level. Gene, transcript and exon annotation data are, therefore, integrated with experimental result data and other well-established databases to setup the primary database of the data-warehouse architecture (Staging Area). Specific DataMarts aggregating data stored in the main Staging Area are then built to navigate and analyze this huge quantity of data. Finally, BI tools such as multidimensional queries, graphics and reports are used to simplify the biological interpretation of Exon Array results.

Results: At this stage of development the Staging Area contains data from publicly available Exon Array experiments. In particular, it stores normalized intensities (at gene and exon level) from experiments carried out on eleven different human tissues and some normal and tumour colon samples. In addition user-provided data may be uploaded in the platform in order to analyze and visualize experimental results in an alternative splicing “perspective”.

An heat map visualization of the hybridization signal at gene, transcript or exon level is provided in all the conditions sampled by experiments stored or uploaded by the user. In this way the user can retrieve and visualize the differential expression of alternative isoforms collected in ASPicDB, in different conditions, using also GeneOntology terms, biochemical pathways (KEGG) and other functional annotation as search criteria.

Abstract 76
- EasyCluster: a fast and efficient gene-oriented clustering tool for large-scale transcriptome data -

Picardi Ernesto[1], Pesole Graziano[1,2]

- [1]Dipartimento di Biochimica e Biologia Molecolare “E.Quagliariello”, Università di Bari, Bari, Italy - [2]Istituto Tecnologie Biomediche, Consiglio Nazionale delle Ricerche, Bari, Italy

Motivation: Expressed sequence tags and full-length cDNAs represent an invaluable source of evidence for inferring reliable gene structures and discovering potential alternative splicing events. In newly sequenced genomes, these tasks could not be easily feasible for several limitations such as the lack of appropriate training sets. However, when expression data (mainly ESTs and FL-cDNAs) are available, they can be consistently used to build EST clusters related to specific genomic transcribed loci. Common strategies employed in the last years, including procedures implemented in UNIGENE, TIGR Gene Index and STACK, are based on sequence similarity mainly detected by means of BLAST, Blat or d2_cluster programs. Moreover, genomic sequences are not always taken into account during the cluster reconstruction. As a consequence, EST sequences can be erroneously grouped leading to inconsistent results when merged in consensus transcripts. In order to improve existing procedures for cluster building and facilitate all downstream annotation analyses, we developed a simple but efficient system to generate gene-oriented clusters (Unigene-like) of ESTs every time a genomic sequence and a pool of related expressed sequences are provided. Our procedure, named here EasyCluster, takes into account the spliced nature of ESTs after an ad hoc genomic mapping. EasyCluster has proved to be highly reliable after a benchmark on a manually curated set of human EST clusters. Therefore, it can be suitably applied for clustering EST in the case a genome sequence is available but not a reliable gene annotation.

Methods: Differently from other existing clustering methods based on similarity searches, EasyCluster takes advantage from the well-known EST-to-genome mapping program GMAP (Wu and Watanabe. Bioinformatics 2005; 21,1859-1875). The main benefit of using GMAP is that it can perform a very quick mapping of whatever expressed sequence onto a genomic sequence attended by an alignment optimization. In particular, GMAP can detect splicing sites according to a so defined “sandwich” dynamic programming that is organism independent.

Given a genomic sequence and a pool of ESTs/FL-cDNAs, EasyCluster first builds a GMAP database of the genomic sequence to speed-up the mapping and a local EST database storing all provided expressed sequences. Subsequently, it runs GMAP program and parses results in order to create an initial collection of pseudo-clusters. Each pseudo-cluster is obtained by grouping ESTs according to the overlap of their genomic coordinates on the same strand. In the next step, EasyCluster refines the EST grouping by running again GMAP on every pseudo-cluster and by including in a cluster only expressed sequences sharing at least one splice site. Finally, for each generated cluster EasyCluster produces a graphical representation in pure HTML code for a simple inspection of results by eyes. EasyCluster is written in python programming language and works on all unix-based platforms where GMAP can be installed.

Results: The EasyCluster program provides EST/FL-cDNA clusters ready to be used in gene prediction pipelines and to detect alternative splicing events. In order to investigate the reliability of EasyCluster we tried to group 256 spliced ESTs related to eleven human homeobox genes (family HOXA) located in the chromosome 7. The same pool of ESTs was used as input in wcd, a new and computationally efficient program to build EST clusters based on sequence similarity (Hazelhurst et al. Bioinformatics. 2008; 13,1542-1546).

EasyCluster was able to reconstruct eleven clusters corresponding to each homeobox gene. In contrast, wcd predicted only nine groups where two of them were related to more than a gene. In particular, ESTs supporting HOXA3 and HOXA4 genes and HOXA9 and HOXA10 genes were clustered together. Our simple results, therefore, demonstrate the reliability of EasyCluster and, in general, of genome-based EST clustering programs over widespread systems based on sequence similarity.

Given the simplicity, flexibility and portability of our system, we are planning to introduce EasyCluster in a more complex pipeline to facilitate the genome-wide detection of the alternative splicing in newly sequenced genomes.

Abstract 77
- GNPIS, THE PLANT INFORMATION SYSTEM OF INRA URGI BIOINFORMATICS PLATFORM -

Steinbach Delphine*[1], Alaux Michael[1], Kimmel Erik[2], Durand Sophie[2], Pommier Cyril[2], Luyten Isabelle[2], Mohellibi Nacer[2], Verdelet Daphne[2], Quesneville Hadi[2]

- [1]Bioinformatics ~ Versailles ~ France - [2]INRA URGI bioinformatics ~ Versailles ~ France

Motivation: URGI (Unité de Recherche Génomique-Info) is an INRA bioinformatics unit dedicated to plants and pest genomics. Created in 2002, one of its mission is to develop and host a genomic and genetic information system called GnpIS, for INRA plants of agronomical interest and their bioagressors. It hosts a bioinformatics platform which belongs to the ReNaBi network and is labelled RIO/IBISA 2007. The URGI maintains an efficient computing environment and offers services covering database conception, software engineering, and bioinformatics. Since 2007, the unit hosts a research team which work is focused on repeats detection and analysis (REPER pipeline). The work presented here will show the description of the actual information system, its development and modular evolution during time, data results since 2000 and details concerning the state of the databases interoperability project.

Methods: GnpIS: information system:
The URGI information system called GnpIS, is a web based system composed of several applications (in Java and Perl) built above a relational database that includes integrated schemas for sequence data, annotation data, mapping data, expression data, proteomic data and SNP data. Since 2005, a new module concerning genetic resource data (SiReGal) was added to the system. Data are submitted by the laboratories through an automatic Web submission tool which allows the checking and the data bulk loading. Web interfaces allow the biologists to query and visualize the data and navigate through them. The ongoing developments are the creation of an interoperability between the genomic and genetic databases modules to allow integrated queries involving all kind of data together and also to be able to skip from one thematic to the other transparently (GnpGenome, GnpSnp, GnpMap, GnpSeq). 2 technologies are in test (in 2008), Biomart (EBI) and Hibernate/Lucene technology, JAVA J2EE technology.
Aster is the key and central module of the information system. It allows the interoperability between the modules.

Results: For all the resources, the databases are available for query either on a public access (http://urgi.versailles.inra.fr), either with an account for partners before publication.

References: See: http://urgi.versailles.inra.fr/about/publications
Grants: Genoplante, ANR Genoplante and INRA

Abstract 78
- A BIOINFORMATICS KNOWLEDGE DISCOVERY APPLICATION FOR GRID COMPUTING -

Marcello Castellano[1], Giuseppe Mastronardi[1], Roberto Bellotti[2], Giacinto Decataldo[1], Luca Pisciotta[1], Gianfranco Tarricone[1]

- [1]DEE – Dipartimento di Elettrotecnica ed Elettronica Politecnico di Bari ~ Bari ~ Italy - [2]Istituto Nazionale di Fisica Nucleare Sezione di Bari e Dipartimento Interateneo di Fisica “M.Merlin” ~ Bari ~ Italy

Motivation: A fundamental activity in biomedical scientific research concerns the Knowledge Discovery process by large amount of biomedical information as documents and data. To have biomedical knowledge baggage more and more updated results a competitive advantage in the scientific progress and a best awareness in biomedical decision support. On the other hand, high performance computational infrastructures, such as Grid technologies, are emerging as available infrastructure to tackle, in principal, the problem of the intensive use of the Information and Communication resources in life science. However, from the application point of view, commodity software solutions are considering to investigate new biomedical information and the use of them is required by the end user. To exploit on a large scale the commodity software for obtain an ICT resources intensive use, a software adapter must be designed.

Methods: The method is based on a layered system software architecture to adapt the no grid based application in a distributed environment. In particular it is a solution to explane how to transform data intensive applications in Single Instruction Multidata stream applications with the aim to enhance bioinformatic application performance. We present a software prototype that interposes between the user’s applications and grid middleware to the purpose to enable distribution of workload between grid nodes. The system, written in JAVA on the Globus Toolkit 4 grid middleware in a GNU/Linux computer grid, includes a graphical user interface, in order to access to a node research system, a load balancing system and a transfer optimizer to reduce communication costs. It has a modular structure to allow the end users to integrate and manage their applications. The prototype does not require that applications be written in a specific language or use specific libraries, it can be used with existing applications after a simple structure standardization. The load balancer analyzes the input data set and the selected computational nodes in order to provide a peer distribution of the workload on the grid. Thus, the transfer optimizer makes a compression of the data set and instruction set sending the compressed data to remote nodes. Finally, the grid computation starts.

Results: We present in details our case study, that is to say, a research of new evidences in biomedical field through a textual classification in terms of symptoms and pathologies recognition through a grid-based system on PubMed documents. Our application starts from 5000 medical scientific publications. It extracts all the possible symptoms and pathologies through Text Mining rules performed by GATE 4.0 using the software adapter here presented for computational grid environment. Then, it creates a number of associative rules that tie, in a probabilistic way, a series of symptoms to one or more pathologies using WEKA4WS grid based software. Finally, we evaluate the contribution in terms of time offered by the grid for the our biomedical application. The speedup factor refers to how much a parallel algorithm is faster than a corresponding sequential algorithm using grid. This factor has been determined for a node number from 1 to 30 for various different size documents. By graphs obtained it is noticed that the use of a grid system offers a profit in this application, this profit is more considerable in correspondence to the increase size document.