Abstract Book

International Conference on Bioinformatics and Computational Biology - BIOCOMP BG 2012

20-21 September, Varna, Bulgaria

Conference venue: Sunny Day Black Sea resort
20-22 September, Varna, Bulgaria

Organized by FP7 REGPOT Project "BioSupport"
University of Plovdiv, Department of Plant Physiology & Molecular Biology

Organizing Committee:

Prof. Ivan Minkov, University of Plovdiv
e-mail: minkov@uni-plovdiv.bg

Dr. Valentina Toneva, University of Plovdiv
e-mail: valton@uni-plovdiv.bg

Dr. Vesselin Baev, University of Plovdiv
e-mail: vebaev@gmail.com

Dr. Elena Aposolova, University of Plovdiv
e-mail: eapostolova@uni-plovdiv.bg
Information:

Sunny Day
9006 Varna, Bulgaria, St. Constantine and Elena
Antonia Mievska Marketing & Sales, Sunny Day Co
Tel. +359 52 361143, Fax: +359 52 361376, Mobile: +359 888843638
E-mail: sunnyday@sunnydaybg.com;
www.sunnydaybg.com

Varna airport
Flight information: Tel. +359 52 573323 Security: Tel. +359 52 573389 Medical care:
Tel. +359 52 573472, +359 52 573467

Triumf taxi: Telephones: +359 52 644444; +359 882 644444; +359 879 644444; +359 898 644444

Cover design: Viktor Ivanov
SPONSORS AND EXHIBITORS OF THE CONFERENCE:

*Platinum sponsors:*

![Platinum sponsors logos]

Supported by:

![Supported by logos]
# Conference Agenda - BIOCOMP BG 2012

## Wednesday, September 19, 2012

<table>
<thead>
<tr>
<th>Time</th>
<th>Event</th>
</tr>
</thead>
<tbody>
<tr>
<td>4:00 pm to 7:00 pm</td>
<td>Arrival, Poster Setup and Registration – Hotel Palace</td>
</tr>
</tbody>
</table>

## Thursday, September 20, 2012

<table>
<thead>
<tr>
<th>Time</th>
<th>Event</th>
</tr>
</thead>
<tbody>
<tr>
<td>8:00 am to 9:00 am</td>
<td>Registration. Posters Setup. Congress center/Hotel Palace</td>
</tr>
</tbody>
</table>

### Morning Session

**Chair: Klaas Vandepoele**

<table>
<thead>
<tr>
<th>Time</th>
<th>Event</th>
</tr>
</thead>
<tbody>
<tr>
<td>9:00 am to 9:15 am</td>
<td>Opening and welcome message.</td>
</tr>
<tr>
<td></td>
<td><strong>Prof. Ivan Minkov, University of Plovdiv, Coordinator of Biosupport project;</strong></td>
</tr>
<tr>
<td></td>
<td><strong>Dr. Vesselin Baev, University of Plovdiv, Group leader Bioinformatics</strong></td>
</tr>
<tr>
<td>9:15 am to 10:15 am</td>
<td>Metagenomics: Next Generation Frontier.</td>
</tr>
<tr>
<td></td>
<td><strong>Keynote speaker: Dr. Jose R. Valverde - Spanish National Center of Biotechnology, Spanish EMBNet node, Spain</strong></td>
</tr>
<tr>
<td>10:15 am to 10:45 am</td>
<td>An approach to a metagenomic data processing workflow</td>
</tr>
<tr>
<td></td>
<td><strong>Dr. Dimitar Vassilev - Agrobioinstitute (ABI), Sofia, Bulgaria</strong></td>
</tr>
<tr>
<td>10:45 am to 11:15 am</td>
<td>Coffee Time 😊</td>
</tr>
<tr>
<td>11:15 am to 11:45 am</td>
<td>FCA as A MEANS for consolidation of clustering results derived from multiple experiments.</td>
</tr>
<tr>
<td></td>
<td><strong>Dr. Veselka Boeva - Technical University, Plovdiv, Bulgaria</strong></td>
</tr>
<tr>
<td>11:45 am to 12:15 pm</td>
<td>Identification of microRNA targets in tomato fruit development using high-throughput sequencing and degradome analysis.</td>
</tr>
<tr>
<td></td>
<td><strong>Dr. Ruud A. de Maagd - Centre for BioSystems Genomics (CBSG), 6700 AB Wageningen, the Netherlands</strong></td>
</tr>
<tr>
<td>12:15 pm to 12:30 pm</td>
<td>Genome wide technologies for detection and analysis of transcription start sites in <em>E. coli</em>.</td>
</tr>
<tr>
<td></td>
<td><strong>Dr. Enrique Morett - Universidad Nacional Autonoma de Mexico, Mexico</strong></td>
</tr>
<tr>
<td>12:30 pm to 12:45 pm</td>
<td>Assembly and annotation of the genomes of Folsomia candida and Orchesella cincta and comparative genomics of stress defense systems.</td>
</tr>
<tr>
<td></td>
<td><strong>Anna Faddeeva-Vakhrusheva - VU Amsterdam, the Netherlands</strong></td>
</tr>
<tr>
<td>12:45 pm to 2:00 pm</td>
<td>Lunch</td>
</tr>
</tbody>
</table>

Lunch
<table>
<thead>
<tr>
<th>Time</th>
<th>Event</th>
</tr>
</thead>
<tbody>
<tr>
<td>2:00 pm - 2:40 pm</td>
<td>Plant and Pathogen Genomics: NGS, Applications and Cross talks.</td>
</tr>
<tr>
<td></td>
<td><strong>Keynote speaker:</strong> Dr. Gaurav Sablok - Research and Innovation Center, Istituto Agrario San Michele (IASMA), Italy</td>
</tr>
<tr>
<td>2:40 pm - 3:00 pm</td>
<td>Approach to <em>in silico</em> filtering of primary SNP-call sets in bacterial genome.</td>
</tr>
<tr>
<td></td>
<td><strong>Nicholas Costa Barroso Lima</strong>– LNCC, Labinfo, Brazil</td>
</tr>
<tr>
<td>3:00 pm - 3:20 pm</td>
<td>The Human Gene Connectome: A Map of Short Cuts for Morbid Allele Discovery.</td>
</tr>
<tr>
<td></td>
<td><strong>Dr. Yuval Itan</strong> - The Rockefeller University, USA</td>
</tr>
<tr>
<td>3:20 pm - 4:00 pm</td>
<td>A computational study on the unusual GW/WG Argonaute binding domain.</td>
</tr>
<tr>
<td></td>
<td><strong>Keynote speaker:</strong> Prof. Wojciech Karłowski - Adam Mickiewicz University, Poland</td>
</tr>
<tr>
<td>4:00 pm - 4:20 pm</td>
<td>Coffee Time 😊</td>
</tr>
<tr>
<td>4:20 pm - 4:40 pm</td>
<td>Illumina presentation (Elta90).</td>
</tr>
<tr>
<td></td>
<td><strong>Theodor Zamfirov</strong> - Elta90 (Illumina), Sofia, Bulgaria</td>
</tr>
<tr>
<td>4:40 pm - 5:00 pm</td>
<td>Identification of miRNAs and their isomiR sequences in human uterine leiomyoma.</td>
</tr>
<tr>
<td></td>
<td><strong>Ivan Milev</strong> - University of Plovdiv, Bulgaria</td>
</tr>
<tr>
<td>5:00 pm - 6:00 pm</td>
<td>Poster Session.</td>
</tr>
<tr>
<td></td>
<td>Go and See the Amazing Posters 😊</td>
</tr>
<tr>
<td>7:30 pm</td>
<td>Gala dinner (Hotel Palace)</td>
</tr>
</tbody>
</table>

**Friday, September 21, 2012**

<table>
<thead>
<tr>
<th>Time</th>
<th>Event</th>
</tr>
</thead>
<tbody>
<tr>
<td>8:00 am - 9:00 am</td>
<td>Registration. Posters set up. Congress center</td>
</tr>
</tbody>
</table>

**Morning Session**

<table>
<thead>
<tr>
<th>Time</th>
<th>Event</th>
</tr>
</thead>
<tbody>
<tr>
<td>9:00 am - 10:00 am</td>
<td>Inferring gene functions and regulatory interactions through integrative genomics in plants.</td>
</tr>
<tr>
<td></td>
<td><strong>Keynote speaker:</strong> Prof. Klaas Vandepoele - VIB / Ghent University, Belgium</td>
</tr>
<tr>
<td>10:00 am - 10:25 am</td>
<td>Evolutionary age-structured cells dynamics models</td>
</tr>
<tr>
<td></td>
<td><strong>Prof. Vitalii Akimenko</strong> - Taras Shevchenko National University of Kyiv, Ukraine</td>
</tr>
<tr>
<td>10:25 am - 10:45 am</td>
<td>A computational model of translation.</td>
</tr>
<tr>
<td></td>
<td><strong>Dr. Dominique Chu</strong> - University of Kent, UK</td>
</tr>
<tr>
<td>Time</td>
<td>Event</td>
</tr>
<tr>
<td>--------------------</td>
<td>-------------------------------------------------------------------------------------------------</td>
</tr>
<tr>
<td>10:45 am to 11:10 am</td>
<td>Coffee Time ☕️</td>
</tr>
</tbody>
</table>
| 11:10 am to 12:00 am | On the quest for Natural Selection: mutations, coadaptation and the emergence of novel functions.  
**Keynote speaker: Dr. Mario Fares - University of Dublin, Ireland** |
| 12:00 am to 12:25 pm | Modeling the Structure of Human Interferon Gamma C-Termini.  
**Dr. Elena Lilkova - University of Sofia “St. Kliment Ohridski”, Bulgaria** |
| 12:25 am to 12:50 pm | An agent-based simulation framework for complex systems.  
**Dr. Gianfranco Politano - Politecnico di Torino, Control and Computer Engineering, Italy** |
| 12:50 pm to 2:00 pm | Lunch                                                                                           |
| **Afternoon Session** | Chair: Jose R. Valverde                                                                            |
| 2:00 pm to 3:00 pm | Structural Bioinformatics: Focus on Proteins with Tandem Repeats.  
**Keynote speaker: Dr. Andrey Kajava - CRBM - Macromolecular Biochemistry Research Center, France** |
| 3:00 pm to 3:20 pm | The exploitation of structural and interaction data to predict protein-ligand binding affinity using machine learning.  
**Dr. Pedro Ballester - EMBL-EBI Wellcome Trust Genome Campus, UK** |
| 3:20 pm to 3:40 pm | CASoX - a software tool to visualize and analyze cavities and clefts in protein structures.  
**Dr. Georg Steinkellner - ACIB - Austrian Centre of Industrial Biotechnology, Austria** |
| 3:40 pm to 4:00 pm | Prediction of intra-communication pathways within the second PDZ domain of PTP1e.  
**Dr. Elisa Cilia - Universite Libre de Bruxelles, Belgium** |
| 4:00 pm to 4:20 pm | Coffee Time ☕️                                                                                   |
| 4:20 pm to 4:35 pm | SureSelect and HaloPlex – NGS Target Enrichment Solutions from Agilent Technologies.  
**Nadia Dimitrova - Diamed Ltd. - Agilent Technologies** |
| 4:35 pm to 5:00 pm | Genomic scale identification of exonic-microRNAs associated with stress responses and development in hexaploid wheat. **Dr. Abdoulaye Baniré Diallo - Université du Québec à Montréal, Canada** |
| 5:00 pm to 6:00 pm | Poster Session.  
Go and See the Amazing Posters ☑️ |
ORAL PRESENTATIONS
Keynote presentation: Dr. Jose R. Valverde
Metagenomics: Next Generation Frontier

José R. Valverde

Centro Nacional de Biotecnología, CSIC, Spain

Abstract: Metagenomics, the collection and analysis of genetic material without previous sample growth in vitro is quickly becoming the next frontier of NGS. Here, we will review some current trends and future challenges being tackled with Metagenomics as well as some of the predictable applications of this exciting new technology, giving due consideration to the main problems to be addressed in any metagenomic analysis. In doing so, we will need to address the whole process, from the early planning stages, through the experimental and analytical phases and up to the final interpretation of results.
Our work on biodiversity analysis will be used as the baseline reference to understand the main problems faced by Metagenomics, and how we have addressed the development of usefully realistic analysis protocols. In addition, we will introduce some new problems we are currently addressing and the intrinsic challenges they provide at the technological level (analysis of massive amounts of new Illumina data), biological level (development of antibiotic resistance) and the analytical level (dealing with repetitive sequence elements).
An approach to a metagenomic data processing workflow

Milko Krachunov\textsuperscript{1}, Dimitar Vassilev\textsuperscript{2}

\textsuperscript{1} Faculty of Mathematics and Informatics, Sofia University, "St. Kliment Ohridski", Sofia, Bulgaria, \textsuperscript{2} Bioinformatics group, AgroBioInstitute, Sofia, Bulgaria

Abstract: Metagenomics is a rapidly growing field that focuses on the study of genetic material collected from largely unexplored heterogeneous biological environments, which has been greatly driven by the ongoing advancements in high-throughput sequencing technologies. As a result, both the data preparation and the subsequent in silico experiments pose unsolved technical and theoretical challenges, as there aren't any well-established approaches, and new expertise and software are constantly emerging.

Due to the nature of the data obtained – which presently lacks any inherent reference points or a standard for validation – every study involves the computational challenges associated with de novo sequencing, which are further exacerbated by the need to deal with a larger degree of uncertainty, significantly larger data and the need to adapt the data processing to your particular experiment, often multiple times.

Our project initial goal is the development and the examination of an approach for error detection and correction aiming to reduce some of the difficulty caused by the lack of references. It relies on the calculation of weighted frequencies based on the local similarity between sequence pairs. In a metagenomic sample, you can't avoid the appearance of related organisms that inevitably lead to the introduction of noise in the data. A carefully defined noise filter can reduce their influence and or even take advantage of their presence as a confirming factor for very well conserved orthologous regions.

During the course of the work it became clear that a tool or a library for managing, running and distributing the processing components would greatly reduce the amount of manual work required to run any metagenomic experiment, and the error detection became a part of a larger project to create a fully-asynchronous library for performing genomic processing, capable of running arbitrary external tools, and providing a general command-line tool for launching preconfigured workflows.

By using an asynchronous network framework and creating interfaces fully adhering to its idioms, the tasks of scheduling, parallelising and distributing the jobs across networked computers becomes easy. It also becomes easier to run a variety of components in variety of combinations while doing basic resource management. Such approach would also facilitate the running and comparison of different error detection approaches with different data processing layers.
FCA as a means for consolidation of clustering results derived from multiple experiments

Veselka Boeva, Anna Hristoskova, Elena Tsiporkova, Elena Kostadinova

1) Computer Systems and Technologies Department, Technical University of Sofia-branch Plovdiv Tsanko Dyustabanov 25, 4000 Plovdiv, Bulgaria vboeva@tu-plovdiv.bg, 2) Department of Information Technology Ghent University - IBBT Gaston Crommenlaan 8 (201), 9050 Ghent, Belgium anna.hristoskova@intec.UGent.be, 3) Sirris, ICT and Software Engineering group The Collective Center for the Belgian technological industry Brussels, Belgium elena.tsiporkova@sirris.be , 4) Computer Systems and Technologies Department Technical University of Sofia-branch Plovdiv Tsanko Dyustabanov 25, 4000 Plovdiv, Bulgaria elli@tu-plovdiv.bg

Abstract: Gene clustering is one of the most important top-down microarray analysis techniques when it comes to extracting meaningful information from gene expression profiles. Clustering algorithms are used to divide genes into groups according to the degree of their expression similarity. Such a grouping indicates that the respective genes are correlated and/or co-regulated, and subsequently indicates that the genes could possibly share a common biological role.

Presently, with the increasing number and complexity of the available gene expression data sets the combination of data from multiple microarray studies addressing a similar biological question is gaining high importance. The integration and evaluation of multiple datasets yields more reliable and robust results since they are based on a larger number of samples and the effects of the individual study-specific biases are diminished. One useful way for integration analysis of the data from different experiments is to aggregate their clustering results into a consensus clustering which both emphasizes the common organization in all the datasets and at the same time reveals the significant differences among them.

In this work, we examine and demonstrate the potential of Formal Concept Analysis (FCA) for consolidation and analysis of clustering results derived separately from a set of microarray experiments studying the same biological phenomenon. We consider two approaches to consensus clustering of gene expression data across multiple experiments. The first algorithm consists of two distinctive steps: 1) a preliminary selected clustering algorithm (e.g. k-means) is initially applied to each experiment separately, which produces a list of different clustering solutions, one per experiment; 2) these clustering solutions are further transformed into a single clustering result by employing FCA, which allows to analyze and extract valuable insights from the data. In the second algorithm, the available microarray experiments are initially divided into groups of related datasets with respect to a predefined criterion (e.g. experimental settings). The rationale behind this is that if experiments are closely related to one another, then these experiments may
produce more accurate and robust clustering solution. Subsequently, the Particle Swarm Optimization (PSO)-based clustering algorithm is applied to each group of experiments separately. The result is a list of different clustering solutions, one for each group. These clustering solutions are pooled together and again analyzed by employing FCA. Notice that FCA produces a concept lattice where each concept represents a subset of genes that belong to a number of clusters. In both algorithms the generated concepts compose the final clustering partition.

The foregoing clustering approaches have been demonstrated to have certain advantages with respect to the traditional consensus clustering techniques, namely both methods: 1) use all data by allowing potentially each experiment (or group of related experiments) to have a different set of genes, i.e. the total set of studied genes is not restricted to those contained into all datasets; 2) are better tuned to each experimental condition by identifying the initial number of clusters for each experiment (or group of related experiments) separately depending on the number, composition and quality of the gene profiles; 3) avoid the problem with ties (i.e. a case when a gene is randomly assigned to a cluster because it belongs to more than one cluster) by employing FCA in order to analyze together all the partitioning results and find the final clustering solution representative of the whole experimental compendium.
Identification of microRNA targets in tomato fruit development using high-throughput sequencing and degradome analysis

Rumyana Karlova, Jan C. van Haarst, Henri van de Geest, Arnaud G. Bovy, Michiel Lammers, Gerco C. Angenent, and Ruud A. de Maagd

Business Unit Bioscience, Plant Research International, 6700 AP Wageningen, The Netherlands, and Centre for BioSystems Genomics (CBSG), 6700 AB Wageningen, the Netherlands

Abstract: MicroRNAs play important roles in plant development through regulation of gene expression by mRNA degradation or translational inhibition. Although tomato (Solanum lycopersicum) is the model system for studying fleshy fruit development and ripening, only a few miRNA targets have been experimentally validated in this species. Thus the role of miRNA action in fruit development and ripening remains largely unknown. Here we present a first step in a more comprehensive study of miRNAs in these processes, by using parallel analysis of RNA ends (PARE) for global identification of miRNA targets and comparing four different stages of tomato fruit development. We identified a total of 181 target genes of 44 miRNA families in tomato. Of these, 166 appeared to be new targets. A large part of the identified targets (68) were coding for transcription factors. Auxin response factors, as well as two known ripening regulators, COLORLESS NON-RIPENING (CNR) and APETALA2a (SlAP2a), with developmentally regulated degradation patterns were identified. The levels of intact messenger of both CNR as well as AP2a are actively modulated during ripening, by miR156/157 and miR172, respectively. In this study we show that miRNA guided cleavage of mRNAs is likely to play an important role in tomato fruit development and ripening.
Genome wide technologies for detection and analysis of transcription start sites in bacteria.

E Morett¹, A Labastida¹, L Vega-Alvarado¹,², V Jimenez-Jacinto¹, M Olvera¹, L Olvera¹, A Mendoza¹, S. Davila³, R Grande³, L Collado⁴, A Hernandez¹, G Gonzalez¹, K Juárez¹, V Moreno⁴, C Vargas⁴, E Sanchez³, B Taboada¹,² and J Collado².

¹Instituto de Biotecnologia, Mexico, ²Centro de Ciencias Aplicadas y Desarrollo Tecnologico, ³Centro de Ciencias Genomicas, UNAM, ⁴Winter Genomics Inc.

Abstract: We are studying gene regulation at a global scale in Escherichia coli and in Geobacter sulfurreducens. Our aims are to map as many transcription start sites (TSSs) and transcriptional units (TU) as possible to get information about promoter and regulatory sites, as well as operon structure. To this end we have generated cDNA libraries from these organisms using different approaches. We started with pyrosequencing, with the Roche’s 454 technology, and detected more than two thousand 5’ end RNA molecules (Men doza et al, PloS ONE, 2009). Recently, we generated cDNA libraries to detect both mono and triphosphate 5’ end RNAs and sequenced them with Illumina GAIIx instrument. For that purpose we made use of a specific primer adapter and a nuclease digestion strategy to eliminate 5’ monophosphate ends. Comparison of the results of 22 different libraries allowed us to discriminate between degradation products (5’monophosphate) from real TSSs (5’triphosphate). Due to the biological and data complexity, these analyses are challenging and require the development of custom algorithms. Critical steps involve 1) maximizing the number of reads that can be used without introducing false alignments, 2) removing biological noise: random transcription and degradation products, 3) identifying the TSSs, 4) visualizing global TSS patterns, and 5) identifying TUs. Condensing the analyses tools into an R/Bioconductor package will guarantee the reproducibility of the work. Results indicate that transcription initiation events are more widespread than anticipated, and that there are large numbers of transcription events within genes and operons and also some in antisense orientation.
Assembly and annotation of the genomes of Folsomia candida and Orchesella cincta and comparative genomics of stress defense systems


*Vrije Universiteit Amsterdam, Institute of Ecological Science, Department of Animal Ecology, De Boelelaan 1085, 1081 HV Amsterdam, the Netherlands ** VU University Medical Center, De Boelelaan 1117, 1081 HV Amsterdam, The Netherlands *** LUMC, Leiden Genome and Technology Center, Eindhovenweg 20, 2333 ZC, Leiden, The Netherlands **** Radboud Universiteit, Comeniuslaan 4, 6525 HP, Nijmegen, The Netherlands ***** Keygene NV, Agro Business Park 9, 6708 PW, Wageningen, The Netherlands

Abstract: Organisms can tolerate to a certain contamination of the ecological niche they inhabit by stress response, which involve induction of stress combating proteins. Some of these proteins have very long evolutionary history while others are specific to the phylogenetic lineage or even the organism. In Animal Ecology Department at the Vrije University Amsterdam we develop classifiers based on the Folsomia candida (F.candida) and Orchesella cincta (O.cincta) gene expression for eco-toxicological testing on specific soil polluters. F.candida was chosen as a model organism for soil quality control and O.cincta as a model organism for studying long term pollution of the soil. In this project we will analyze the F. candida and O.cincta genome and compare them to other arthropods to reveal signatures of evolution of stress-responses and explain the phenotypic plasticity of these organisms in response to contamination. Combination of Illumina and 454 next-generation sequencing (NGS) technologies will be implemented to obtain complete genome and transcriptome of F.candida and O.cincta. In this context we collaborate with Leiden Genome Technology Center (Prof. Dr. J. den Dunnen). During the project we are going to assemble the complete genome and transcriptome of F.candida and O. cincta by using several available assembly platforms (CLC, AbySS). Subsequently, gene annotation and phylogenetic analysis will allow us to further characterize the genome content. We hypothesize, that occupation of different niches influenced the evolution of stress defense systems. Not only evolution of genes families will be investigated, but also cis-regulatory motives among stress-response gene networks will be analyzed to study plasticity of stress response pathways.
Keynote presentation: Dr. Gaurav Sablok
Plant and Pathogen Genomics: NGS, Stress Adaptation and Cross talks
Sustainable Agro-ecosystems and Bioresources Department,

Dr. Gaurav Sablok

IASMA Research and Innovation Centre,
Fondazione Edmund Mach, Via E. Mach 1,
38010 San Michele all’Adige, (TN), Italy

Abstract: Plant disease epidemics and stress adaptation genetics, are two serious global questions to be addressed in this era of global warming. The host-pathogen adaptation question becomes more imperative, when it is centered towards the host-pathogen relationship and its macro and micro-evolutionary concerns. Next generation sequencing technologies has opened the gateway for a wide visualization of the interactions between the host and pathogen. Isobaric tagging in quantitative mass spectrometry (iTRAQ) has played a major role in the mapping of the interactome during the time course of infection. The talk will be focused on the utilization of the Next Generation Sequencing technology to explore the adaptation of the fungal and the bacterial pathogens to the host. One section of the talk will be focused on the adaptation of the plants to the stress on the basis of the genome wide bias patterns and will shed some light on the evolutionary perspective of the genome adaptation to the stress. The talk will mark the release of a Organelle genome repeat database (ChloroMitoSSRDB) and its role in evolutionary genomics.
Approach to in silico filtering of primary SNP-call sets in bacterial genome

Nicholas Costa Barroso Lima, Maurício Egídio Cantão, Marisa Fabiana Nicolás

LNCC, Brazil

Abstract: Background: Single Nucleotide Polymorphisms (SNPs) are nucleotide substitutions per site between two sequences. Currently, several studies are focusing on the detection of SNPs in bacterial genomes. To analyze data from the NGS, a bioinformatics approach is mandatory to achieve the SNP-call sets. Thus, a pipeline for detection and filtering of primary SNP-call sets in bacterial genome was developed.

Methods: Scripts in PERL were written to function as filters for the primary SNP-call sets. These filters function in three levels returning a more reliable SNP-call set. SNPs were also classified regarding their location on coding sequence (CDS) or intergenic region. The pipeline was used to detect SNPs among four Klebsiella pneumoniae strains. The reference genome was the clinical isolate Kp13 (Ramos et al., BMC Microbiology, 12:173, 2012) and the comparisons were made against two pathogenic strains (MGH 78578 and NTUH-K2044) and a nitrogen fixating strain (342). SNPs between paired comparisons of these strains were further investigated to assess association with virulence and bacterial fitness.

Results: Three primary SNP-call sets were generated from the comparisons Kp13 vs 342, Kp13 vs MGH 78578 and Kp13 vs NTUH- K2044. Then, from these primary SNP-call sets we obtained nine filtered SNP-call sets. SNPs prevalence is almost ten-fold higher in Kp13 vs 342 than those in the other two comparison analyses. This result may be explained by the different lifestyle of 342 (endophytic plant association). By Poisson Probability Distribution are expected 0-4 SNPs per CDS for Kp13 vs 342 comparison and 0-2 SNPs for Kp13 vs NTUH-K2044 and Kp13 vs MGH 78578 comparisons. Interestingly, CDSs in Kp13 with more SNPs than expected by Poisson Probability are associated with plant growth hormones in comparison with nitrogen fixing strain 342, T4SS and siderophore in comparison with NTUH-K2044 and antimicrobial resistance in comparison with MGH 78578.
The Human Gene Connectome: A Map of Short Cuts for Morbid Allele Discovery

Yuval Itan, Shen-Ying Zhang, Guillaume Vogt, Avinash Abhyankar, Lluis Quintana-Murci, Laurent Abel and Jean-Laurent Casanova

The Rockefeller University, USA, INSERM, Pasteur Institute

Abstract: High-throughput genomic data reveal thousands of gene variants per patient and it is often difficult to decipher which of these variants underlie disease in individuals. At the population level however, there can be some level of phenotypic homogeneity, with alterations of specific physiological pathway underlying the pathogenesis of a particular disease. Here we describe the human gene connectome (HGC) as a new approach facilitating the interpretation of abundant genetic data, guiding subsequent experimental investigations. We identify the set of shortest plausible biological distances, routes, and degrees of separation between all pairs of human genes, by applying a shortest distance algorithm to the full human gene network. We demonstrate a hypothesis-driven application of the HGC in which we generate a TLR3-specific connectome, which may be useful for the genetic dissection of herpes simplex virus encephalitis of childhood. We also developed the functional genomic alignment (FGA) approach from the HGC. In FGA, the genes are clustered according to their biological proximity (rather than the traditional evolutionary genetic distance), as estimated from the HGC. The HGC and FGA should facilitate the genome-wide discovery and experimental validation of disease-causing alleles.
Keynote presentation: Prof. Wojciech Karlowski
A computational study on the unusual GW/WG Argonaute binding domain

Wojciech Karlowski

Laboratory of Computational Genomics (http://comgen.pl)
Institute for Molecular Biology and Biotechnology, Poznan, Poland

Abstract: RNA silencing is an essential component of gene expression regulation. Small RNAs (20-30 nt) carry the sequence information that guides the silencing effector complexes to specific targets on DNA (Transcriptional Gene Silencing) or RNA (Posttranscriptional Gene Silencing). Argonaute (AGO) proteins are factors common to all these complexes, but their complete protein composition is not yet sufficiently explored. A recent advancement in this subject is the observation that a novel element of such complexes in eukaryotes contains an unusual, repetitive peptide sequence binding AGO proteins – the GW/WG motif. The WG/GW motifs were first described in the human GW182 protein. Since then, AGO-binding properties have been demonstrated for several WG/GW motif-containing proteins, including GW182 homologs in Caenorhabditis elegans (AIN-1 and AIN-2), Drosophila (GAWKY), Schizosaccharomyces pombe (Tas3) and Arabidopsis thaliana large subunit of RNA PolV (NRPE1).

Strikingly, a comparison of the Arabidopsis NRPE1 GW/WG sequence with sequences of other plants shows little conservation in the repeats other than the WG/GW pairs, even between relatively closely related species. The lack of sequence similarity, however, does not influence the functional universality of the domain. For example, in a domain swapping experiment, the human sequence with GW/WG motifs of GW182 protein can substitute an analogous region in the NRPE1 protein. Such chimeric construct binds to AGO4 in vitro and restores its DNA methylation activity, while the amino acid sequences show barely detectable similarity. The unusual character of the AGO-binding WG/GW domain imposes several challenges and raises important questions. The highly divergent sequence, variable amino acid lengths of the motifs and irregular number of repeats make the detection of the domain very difficult. Therefore, one of the most important tasks is to search for methods that will allow efficient annotation of this type of sequences. Most of the current annotation tools rely on sequence preservation – higher sequence similarity means a conserved function and a higher probability of correct protein description. However, in the case of WG/GW domain, the sequence shows a high level of divergence. To address this issue, we have developed in our laboratory a comprehensive set of bioinformatics tools, which allow reliable identification of the domain, determination of its boundaries as well as statistical quantification of the predictions’ quality.

Recently, research on the function of WG/GW domains has become even more complex and interesting. Several papers have been published that depict a novel
function for the GW-based motifs. It has been demonstrated that C-terminal regions of the GW182 proteins from humans and Drosophila function by recruiting components of the CCR4–NOT complex inducing deadenylation and mRNA decay. Moreover, it has been presented that tryptophan residues located in the C-terminal region of the protein, similar as in AGO-binding activity, are required and sufficient for the interaction with silencing complex.

The newly discovered molecular roles for W-based motifs fill another gap in the description of functional small RNA silencing complexes. At the same time they raise new questions and open new exciting perspectives for creative research, where computational methods play a crucial role in exploring new solutions, seeking new answers. It is clear that the quest for not-yet-identified proteins with AGO or CCR4-NOT binding domains has just begun. The question if two identified functions saturate all the possible activities of W-based domains, remains open. However, it is obvious now that the time has come to focus on the non-conserved, unstructured parts of proteins with more scientific attention.
Version 2 of the Illumina’s desktop Next Generation Sequencing platform - The MiSeq - what is possible?

Theodor Zamfirov

Elta90, Illumina

Abstract: Illumina is the market leader and provider of total solution in Genetic research, from the unique SBS chemistry, Hi Tech sequencing platforms to the data analyses software assuring the best possible quality of readings and highest on the market accuracy above Q30. The presentation will give an overview of the technology, the capacity of the system, new improvements in technical aspects, the chemistry and the clouding storage of data, called BaseSpace and in the end the automatic on board data analyzing softwares.
Identification of miRNAs and their isomiR sequences in human uterine leiomyoma.

I. Milev, B. Georgieva, G. Minkov, V. Baev

University of Plovdiv, Dept. Plant Physiology and Molecular Biology, Bulgaria

High-throughput sequencing or deep sequencing has emerged as a direct small RNA examination method. Profiling of microRNAs by deep sequencing allows the discovery of novel miRNAs, measuring absolute abundance as well as detection of variants from “reference” mature sequences that are called “isomiRs”. Deep sequencing generates millions of small RNA sequence reads from a given sample profiling and provides a framework for exploring the transcriptional complexity. However, high-throughput sequencing data sets present new problems related to data processing and extraction of all information from the huge amount of obtained data.

miRNAs play important roles in cellular control and in various disease states such as cancers. Identifying the whole repertoire of miRNAs and understanding their expression patterns is therefore an important goal.

Uterine leiomyomas are benign neoplasms of the smooth muscle cells of the uterus and are the most common tumors in women. Although they are not malignant, these tumors may cause significant reproductive and gynecological problems, such as menorrhagia, dysmenorrhea, chronic pelvic pain, as well as infertility and pre-term delivery. However, there is little understanding of the processes of leiomyomas growth and regression at the cellular and molecular levels.

Here, we present a new web-based tool for identification of miRNA, their cognate isomiRs and investigation of differentially expressed miRNAs using NGS datasets. The software also detects isomiRs of a miRNA with higher copy number relative to their mature reference sequence indexed in mirBase.
Keynote presentation: Klaas Vandepoele
Inferring gene functions and regulatory interactions through integrative genomics in plants

Prof. Klaas Vandepoele

VIB / Ghent University, Belgium

A major challenge is to unravel how genes interact and are regulated to exert specific biological functions. One of the major challenges of computational biology is the integration and exploitation of genome-wide data sets such as transcriptome and interactome data, metabolomics and other -omics data, and large-scale phenotyping. Data integration is often performed through gene network analysis and the resulting networks can increase our knowledge of functional gene relationships and the interplay of different types of interactions. Depending on the type of interaction data, different types of modules can be defined and examples in Arabidopsis include coexpression, protein complexes, and modules grouping genes that are regulated by the same transcription factor (TF). In this talk I will present how Arabidopsis genes are organized into gene modules based on different experimental and computational data types and discuss the functional and regulatory properties of these modules. Furthermore, I will highlight how, based on large-scale expression compendia grouping multiple experiments, the guilt-by-association principle can be applied in a comparative manner to study conserved modular gene programs, identify cis-regulatory elements, or predict functions of unknown genes in different model plants. Apart from describing the different steps required to systematically compare expression data across plants species, some applications and case studies will be presented.
Evolutionary age-structured cells dynamics models

Vitalii Akimenko, Yurii Zahorodnii.

Taras Shevchenko National University of Kyiv, Ukraine

Abstract: Age-structured models describe the behaviour of differ evolutionary biological systems including the process of cell dividing or birth of living organism. Corresponded researches can be involved in the projects of modelling of growth in physical and biological systems in the cells scale. These models are based on the systems of linear partial differential equations of hyperbolic type – transport equations with special integral boundary conditions. In particular case of linear system for cell density distribution one obtained an analytical solution. This approach allows us to analyse basic lows of behaviour of macroparameters of isolated biological systems and resolve the problems of biological parameters identification and optimal control and get analytical solutions for important applied systems.

A computational model of translation

Dominique Chu and Tobias von der Haar

University of Kent, United Kingdom

Abstract: We present a computational model of translation. This model is genome wide and takes into account all the important steps of translation, including initiation, elongation, ribosome recycling and ribosome-ribosomes interactions on transcripts. In this talk we will first present the software and show how it can be parameterised for any organism. We will then present detailed results obtained for a parameterisation for yeast and also show some experimental results that confirm the predictions of the model.
Keynote presentation: Mario Fares
On the Quest for Natural Selection: Mutations, Coadaptation and the Emergence of Novel Functions

Mario Fares

University of Dublin

Abstract: Evolutionary biologists have long sought finding the origin of life by linking all life forms through phylogenetic trees. This quest has been based on the assumption that a single ancestor for all of life differentiated into multiple descendant species, each adapted by natural selection to one of a new set of ecological niches. But, how does natural selection act? Many decades since the publication of the “Origin of Species” have been devoted to uncover the mechanisms underlying natural selection, in particular using molecular data, yet the discovery of many such mechanisms have only led to new questions. All such efforts have allowed us to conclude that: (a) mutations are the main source of evolution; (b) The evolution of new traits or functions depends on an importantly large set of conditions; and (c) many biological phenomena, such as gene and genome duplication, play a fundamental role in accelerating and mediating the emergence of novel functions and biological complexities. Here I discuss the roles of mutations and their interactions and gene and genome duplications in the emergence of biological novelties.
Modeling the Structure of Human Interferon Gamma C-Termini

Peicho Petkov, Elena Lilkova, Nevena Ilieva, and Leandar Litov

University of Sofia "St. Kliment Ohridski", Bulgaria

Abstract: Human interferon gamma (hIFN-g) is an important signaling molecule, which plays a key role in the formation and modulation of immune response and in gene activity regulation [1]. The protein represents a non-covalent homodimer of 144 amino acids long monomers, organized in six alpha-helices, which are linked by short unstructured regions, followed by a long unstructured C-terminal domain (the last 21 amino acids) [2]. The hIFN-g expresses its activity through the high-affinity interaction with its own species-specific extracellular receptor (hIFNGR). Three regions of the hIFN-g molecule are responsible for receptor binding, one of which is a short sequence in the C-terminal tail (residues 128-131). The hIFN-g C-terminal region is highly positively charged and therefore highly susceptible to proteases. Experimental studies demonstrated that removal of up to 9 residues causes gradual increase in the activity of the protein. However, the complete removal of the flexible C-terminus inactivates the cytokine. Although all studies undoubtedly proved the modulating effect of the unstructured C-terminal region on hIFN-g activity, they failed to explain the molecular mechanism of its action due to the lack of X-ray diffraction data. The extremely controversial conclusions about the function of the hIFN-g C-terminus as well as the lack of a reasonable model explaining its role in the receptor binding prompted us to model its structure by means of molecular dynamics (MD) simulations. This model was then used for investigating the interaction of hIFN-g with hIFNGR1 to reveal the mechanism of this interaction and to shed more light on the role of the C-terminus in hIFN-g -- hIFNGR1 binding.

We performed extensive MD folding simulations (200 ns each), which unambiguously show that the C-termini tend to adopt a compact conformation together with the globular part of the cytokine. The obtained trajectories were subjected to cluster analysis in order to detect the population intensity of the various conformational states. Analysis of the solvation free energy of the centroids of the most populated clusters confirmed their energetic favourability. For validation of the obtained results, the same simulations were performed using different force fields (GROMOS 53a6, AMBER 99SB, and CHARMM 27) with different simulation packages (GROMACS 4.5.4 and NAMD 2.9). Despite the expected minor variations in the details, our conclusions remained unchanged.

Support from Grants DO02-115/15.12.2008, DCVP 02/1/29.12.2009, DID-02-30/2009 and DRG 02/05/2010 of the Bulgarian National Science Fund is acknowledged. Calculations were performed at the Bulgarian Supercomputing Center and a local Linux cluster at the Atomic Physics Department of the Faculty of Physics of University of Sofia “St. Kliment Ohridski”
Using Boolean Networks to Model Post-transcriptional Regulation in Gene Regulatory Networks

A. Benso, S. Di Carlo, G. Politano, A. Savino, H. Ur Rehman

Politecnico di Torino, Italy

Abstract: In Gene Regulatory Networks research there is a considerable lack of techniques and tools to understand these networks from a System Biology point of view. The typical biological approach is to reconstruct a particular network from expression patterns, then try to validate its dynamics by simulation, use simulation to analyze its reactions to some perturbations, and finally go back “in vitro” to validate the simulation results. Nevertheless, when the goal is to understand the high-level, general mechanisms that allow these networks to work or to be stable under mild perturbations, this type of approach has shown very strong limitations. In this work we want to better understand the role of miRNA as a stabilizing mechanism in gene regulatory networks. Boolean networks have been recently used to better understand the structural and dynamical properties of regulatory networks. Attractors and ergodic sets have been easily correlated with many of the typical biological cell behaviors (cancer, differentiation, pluripotential, ...). The most widely used model are nevertheless very simple, and work under too strict constraints.

We are defining an enhanced model based on Boolean Networks but also able to take into account post-transcriptional regulation and possibly be extended to other regulatory mechanisms (e.g. ceRNA) that have been already proven crucial in vivo. The final goal is to try to understand if the wide number of miRNA targets constitutes a structural network-stability mechanism used to make the network immune to “regulatory” noise. To achieve this result we evolve the modified Boolean networks for high or low sensitivity to perturbations, and then analyze the resulting networks to understand if specific structural patterns containing miRNA-like post-transcriptional regulatory elements can be correlated with the network stability.
An agent-based simulation framework for complex systems

A. Benso, S. Di Carlo, G. Politano, A. Savino, H. Ur Rheman

Politecnico di Torino, Italy

Abstract: In this abstract we present a new approach to the simulation of complex systems as biological interaction networks, chemical reactions, ecosystems, etc... It aims at overcoming previously proposed analytical approaches that, because of several computational challenges, could not handle systems of realistic complexity. The proposed model is based on a set of agents interacting through a shared environment. Each agent functions independently from the others, and its behavior is driven only by its current status and the “content” of the surrounding environment. The environment is the only “data repository” and does not store the value of variables, but only their presence and concentration. Each agent performs 3 main functions:
1) it samples the environment at random locations
2) based on the distribution of the sampled data and a proper Transfer Function, it computes the rate at which the output values are generated
3) it writes the output “products” at random locations.
The environment is modeled as a Really Random Access Memory (R2AM). Data is written and sampled at random memory locations. Each memory location represent an atomic sample (a molecule, a chemical compound, a protein, an ion, ...). Presence and concentration of these samples are what constitutes the environment data set. The environment can be sensitive to external stimuli (e.g., pH, Temperature, ...) and can include topological information to allow its partitioning (e.g. between nucleus and cytoplasm in a cell) and the modeling of sample “movements” within the environment.
The proposed approach is easily scalable in both complexity and computational costs. Each module could implement a very simple object as a single chemical reaction or a very complex process as a gene translation into a protein. At the same time, from the hardware point of view, the complexity of the objects implementing a single agent can range from a single software process to a dedicated computer or hardware platform.
Keynote presentation: Dr. Andrey V. Kajava
Structural bioinformatics: focus on proteins with tandem repeats

Andrey V. Kajava

Centre de Recherches de Biochimie Macromoléculaire, CNRS Université Montpellier, France

Abstract: A significant portion of proteins carrying fundamental functions contain arrays of tandem repeats. Furthermore, over the last years a number of evidences has been accumulated about the high incidence of tandem repeats in the virulence proteins of pathogens, toxins and allergens. Genetic instability of these regions can allow a rapid response to environmental changes and thus can lead to emerging infection threats. In addition, the tandem repeats frequently occur in amyloidogenic, prion and other disease-related human sequences. Thus, discovery of these domains and their structure-function study promise to be a fertile direction for research leading to the identification of targets for new medicaments and vaccines. This presentation will provide a survey of current challenges in this area including: identification of protein repeats in genomes, structural classification of proteins with repeats, prediction and molecular modeling of the 3D structure of these proteins, inferring of their functions.
The exploitation of structural and interaction data to predict protein-ligand binding affinity using machine learning

Pedro J Ballester¹, Adrian Schreyer² and Tom L. Blundell²

1 European Bioinformatics Institute, Wellcome Trust Genome Campus, Hinxton - CB10 1SD, U.K., 2 Dept. of Biochemistry, University of Cambridge, 80 Tennis Court Rd, Cambridge - CB2 1GA, U.K.

Abstract: Accurately predicting the binding affinities of large sets of diverse molecules against a range of macromolecular targets is an extremely challenging task. The scoring functions that attempt such computational prediction exploiting structural data are essential for analysing the outputs of molecular docking, which is in turn an important tool for drug discovery, chemical biology and structural biology. Conventional scoring functions assume a predetermined theory-inspired functional form for the relationship between the variables that characterise the X-ray structure of the complex and its predicted binding affinity. The inherent problem of this approach is in the difficulty of explicitly modelling the various contributions of intermolecular interactions to binding affinity.

A new family of scoring functions able to effectively exploit much larger amounts of experimental data and circumvent the need for an explicit functional form has recently been introduced¹. These machine learning regression models have been shown to outperform a wide range of state-of-the-art scoring functions in a publicly available benchmark². Here we will investigate the impact of the way the protein-ligand complex is encoded on the predictive power of the resulting scoring function. This systematic battery of numerical experiments has led to an even more accurate scoring function. An additional contribution of this study is in the critical assessment of modelling assumptions. In the light of our performance results and considering the uncertainty introduced by the static nature of crystal structures, we will discuss the role of interatomic distance cutoffs and binning as well as protonation states for the binding affinity prediction of a protein-ligand complex.

REFERENCES
CASoX - a software tool to visualize and analyze cavities and clefts in protein structures

Georg Steinkellner, Gustav Oberdorfer and Karl Gruber

ACIB GmbH, Graz, Austria; Institute of Molecular Biosciences, University of Graz, Austria

Abstract: We have developed a plug-in for the visualization tool PyMOL for a convenient and flexible representation of cavity information. This will lead to a better and enhanced analysis of cavities in distinct cases and will also give new insights into specific structural behaviors like water channels or exit tunnels. Cavity and cleft information can be very useful to understand biologically or industrially important protein behaviors. The shape of the cavity as well as the size, solvent accessibility and physico-chemical properties are often relevant for binding and conversion of enzyme specific substrates. A convenient calculation, visualization and interaction tool for cavities and clefts will ultimately help to analyze, interpret and present cavity information more easily. Most cavity programs output the best proposed cavities for different prediction purposes which are defined by internal program settings. This makes it more likely to miss some cavity features or to interpret cavity or cleft information differently. With CASoX you can show all calculated cavities and select for yourself which cavity or cavities are important merely by looking to the overall shape and location of the cavities or by playing with the standard settings. Therefore also active sites which are not in a closed cavity can be viewed and analysed. It is easier to identify water or substrate tunnels which are not directly visible because e.g. an amino acid may block the entry (or exit). It is also possible to calculate and map hydrophobicity and accessibility values onto the cavity points. (Figure 1) By utilizing the capabilities of PyMOL every calculated cavity can be viewed and every cavity point is accessible for selection. Presets are provided to use the tool very quickly without going deeper into the advanced features and settings, providing a fast cavity calculation tool to be used as a standard analysis instrument.
Cupin E8WYN5 binding pocket analysis and visualization via CASoX plugin

Andrzej LYSKOWSKI, Georg STEINKELLNER, Kerstin STEINER, Helmut SCHWAB and Karl GRUBER

ACIB GmbH, Graz, Austria

Abstract: The recombinant protein AciX9_0562 from Acidobacterium sp. MP5ACTX9 (UniProt ID E8WYN5) has recently been crystallized and preliminary structure has been calculated. A structure of 131 amino-acid protein of uncharacterized function with a molecular weight of 14.25 kDa revealed a large cavity with a putative active site located around a metal ion bound at the bottom. Here we present a novel cupin E8WYN5 structure and a visualization of binding pockets and the preliminary analysis via a CASoX PyMOL plugin. The graphical interpretation of the cavity volume as well as its biophysical properties are presented in easy to interpret graphical form. The standard setting of the CASoX plugin allow for the rapid visualization of multiple properties including but not limited to cavity proportions and its hydrophobicity. Ability to rapidly visualize the properties is essential to interpret and analyze specific activity of the enzyme particularly in case when the exact native activity is unknown.
Prediction of intra-communication pathways within the second PDZ domain of PTP1e

Elisa Cilia (1), Geerten W. Vuister (2) and Tom Lenaerts (1,3)

(1) MLG, Département d’Informatique, Université Libre de Bruxelles, Boulevard du Triomphe CP212, 1050 Brussels, Belgium (2) Department of Biochemistry, University of Leicester, Henry Wellcome Building, Lancaster Road, Leicester LE1 9HN, United Kingdom (3) AI-lab, Vakgroep Computerwetenschappen, Vrije Universiteit Brussel, Pleinlaan 2, 1050 Brussels, Belgium

Abstract: Experimental NMR relaxation studies have shown that peptide binding induces dynamical changes at the side-chain level throughout the second PDZ domain of PTP1e, identifying as such the collection of residues involved in long-range communication. Even though different computational approaches have identified subsets of residues that were qualitatively comparable, no quantitative analysis of the accuracy of these predictions was thus far determined. We show that our information theoretical method produces quantitatively better results with respect to the experimental data than some of these earlier methods. Moreover, we provide a global network perspective on the effect experienced by the different residues involved in the process. We show furthermore that these predictions are consistent within both the human and mouse variants of this domain. Together, these results improve the understanding of intra-protein communication and allostery in PDZ domains, underlining at the same time the necessity of producing similar data sets for further validation of this kind of methods.
SureSelect and HaloPlex – NGS Target Enrichment Solutions from Agilent Technologies.

Nadia Dimitrova

Diamed Ltd, Agilent Technologies

Agilent’s SureSelect market leading platform provides a complete portfolio of catalog and custom products, providing the flexibility you need, for your discovery to follow-up studies. The new Human All Exon V4 and V4+UTRs kits provide comprehensive coverage of the major genome databases.

New HaloPlex products combine the speed of PCR with the sensitivity of hybridization providing a robust solution for targeting smaller capture regions. By combining single-tube target amplification and removing the need for library preparation, the total sample preparation time and cost is reduced without the need for dedicated instrumentation or automation. The availability of 96 indexes and compatibility with different desktop sequencing and high-throughput platforms provide you the flexibility you need to run your experiments. Finally, the assay has premium performance, with high specificity, uniformity and variant detection.
Genomic scale identification of exonic-microRNAs associated with stress responses and development in hexaploid wheat

Abdoulaye Baniré Diallo

Université du Québec à Montréal, Canada

Abstract: MicroRNAs (miRNAs) are a family of small non-coding RNA that regulate gene expression involved in cellular metabolism, growth, development and stress responses. They are emerging as important regulators of gene expression in eukaryotes including plants. In plants, hundreds of miRNAs were reported in species such as A.Thaliana, M. Truncatula and O. Sativa. However, in wheat (Triticum aestivum), less than 40 miRNAs are present in miRBase with very little known about their functions. Here, we predicted 869 expressed microRNAs and there associated biological function in ten conditions related to cold and vernalization, salt, aluminum, development and tissue specificity. To accurately predict microRNA, we developed a robust pipeline with different cross validation level and a unique machine learning classifier based on 35 important features associated to the true microRNA hairpins. This classifier permits to avoid most of false positive microRNAs predicted by HHMMir and Mipred as well as being less rigid than Mircheck and Meyer’s criteria for plant microRNA prediction. Finally, it is important to notice that different set of the identified microRNAs in different conditions have been found to be enriched for different biological processes according to gene ontology such as response to stress for cold regulated microRNAs, flower development for microRNA regulated in development condition, and others.
POSTER PRESENTATIONS
P1

ChloroMitoSSRDB: Open source repository of perfect and imperfect repeats in organelle genomes for evolutionary genomics

Martina Popova¹, Suresh B. Mudunuri², Sujan Patnana³, Vesselin Baev¹, Gaurav Sablok⁴*

¹Department of Molecular Biology, Bioinformatics Group, University of Plovdiv, Tzar Assen, Plovdiv, Bulgaria; ²Department of Computer Science & Engineering, Vishnu Institute of Technology, Vishnupur, Bhimavaram, Andhra Pradesh, India; ³TalentSprint Edu. Services, International Institute of Information Technology (IIIT) Campus, Gachibowli, Hyderabad, Andhra Pradesh, India; ⁴Sustainable Agro-ecosystems and Bioresources Department, IASMA Research and Innovation Centre, Fondazione Edmund Mach, Via E. Mach 1, 38010 San Michele all’Adige, (TN), Italy

Abstract: Microsatellites are repetitive stretches of nucleotide (A, T, G, C), which are distributed as a single base pair stretches or a combination of two- to six-nucleotides non-randomly with in the coding or the non-coding region of the genome. ChloroMitoSSRDB is a complete curated web-oriented relational database of perfect and imperfect repeats in organelle genomes. The present version of the database contains perfect and imperfect Simple Sequence Repeats (SSRs) for 2,161 organelle genomes (1,982 mitochondrial and 179 chloroplast genomes). We detected a total of 5,838 chloroplast perfect SSRs; 37,297 chloroplast imperfect SSRs; 5,898 mitochondrial perfect SSRs; 50,355 mitochondrial imperfect SSRs across these genomes. The repeats have been further hyperlinked to the annotated gene regions (coding or non-coding) and a link to the corresponding gene record in NCBI (www.ncbi.nlm.nih.gov/) to identify and understand the positional relationship of the repetitive tracts. ChloroMitoSSRDB is connected to a user-friendly web interface, which gives useful information associated with the location of the repeats (coding and non-coding), size of repeat, motif and length polymorphism. ChloroMitoSSRDB will serve as a repository for developing functional markers for molecular phylogenetics, estimating molecular variation across various species.
P2

Genome wide technologies for detection and analysis of transcription start sites in bacteria.

E Morett¹, A Labastida¹, L Vega-Alvarado¹,², V Jimenez-Jacinto¹, M Olvera¹, L Olvera¹, A Mendoza¹, S. Davila¹, R Grande¹, L Collado⁴, A Hernandez¹, G Gonzalez¹, K Juárez¹, V Moreno⁴, C Vargas⁴, E Sanchez³, B Taboada¹,² and J Collado².

¹Instituto de Biotecnologia, Mexico, ²Centro de Ciencias Aplicadas y Desarrollo Tecnologico, ³Centro de Ciencias Genomicas, UNAM, ⁴Winter Genomics Inc.

Abstract: We are studying gene regulation at a global scale in *Escherichia coli* and in *Geobacter sulfurreducens*. Our aims are to map as many transcription start sites (TSSs) and transcriptional units (TUs) as possible to get information about promoter and regulatory sites, as well as operon structure. To this end we have generated cDNA libraries from these organisms using different approaches. We started with pyrosequencing, with the Roche’s 454 technology, and detected more than two thousand 5’end RNA molecules (Mendoza et al, PloS ONE, 2009). Recently, we generated cDNA libraries to detect both mono and triphosphate 5’ end RNAs and sequenced them with Illumina GAIIx instrument. For that purpose we made use of a specific primer adapter and a nuclease digestion strategy to eliminate 5’ monophosphate ends. Comparison of the results of 22 different libraries allowed us to discriminate between degradation products (5’monophosphate) from real TSSs (5’triphasphate). Due to the biological and data complexity, these analyses are challenging and require the development of custom algorithms. Critical steps involve 1) maximizing the number of reads that can be used without introducing false alignments, 2) removing biological noise: random transcription and degradation products, 3) identifying the TSSs, 4) visualizing global TSSs patterns, and 5) identifying TUs. Condensing the analyses tools into an R/Bioconductor package will guarantee the reproducibility of the work. Results indicate that transcription initiation events are more widespread than anticipated, and that there are large numbers of transcription events within genes and operons and also some in antisense orientation.
Assembly and annotation of the genomes of Folsomia candida and Orchesella cincta and comparative genomics of stress defense systems

A. Faddeeva-Vakhrusheva*, D. Sie**, K. Kraaijveld***, J. Mariën*, H. Op den Camp****, E. Datema***** and D. Roelofs*

*Vrije Universiteit Amsterdam, Institute of Ecological Science, Department of Animal Ecology, De Boelelaan 1085, 1081 HV Amsterdam, the Netherlands ** VU University Medical Center, De Boelelaan 1117, 1081 HV Amsterdam, The Netherlands *** LUMC, Leiden Genome and Technology Center, Einthovenweg 20, 2333ZC, Leiden, The Netherlands **** Radboud Universiteit, Comeniuslaan 4, 6525 HP, Nijmegen, The Netherlands ***** Keygene NV, Agro Business Park 9, 6708PW, Wageningen, The Netherlands

Abstract: Organisms can tolerate to a certain contamination of the ecological niche they inhabit by stress response, which involve induction of stress combating proteins. Some of these proteins have very long evolutionary history while others are specific to the phylogenetic lineage or even the organism.

In Animal Ecology Department at the Vrije University Amsterdam we develop classifiers based on the Folsomia candida (F.candida) and Orchesella cincta (O.cincta) gene expression for eco-toxicological testing on specific soil polluters. F.candida was chosen as a model organism for soil quality control and O.cincta as a model organism for studying long term pollution of the soil.

In this project we will analyze the F. candida and O.cincta genome and compare them to other arthropods to reveal signatures of evolution of stress-response and explain the phenotypic plasticity of these organisms in response to contamination.

Combination of Illumina and 454 next-generation sequencing (NGS) technologies will be implemented to obtain complete genome and transcriptome of F.candida and O.cincta. In this context we collaborate with Leiden Genome Technology Center (Prof. Dr. J. den Dunnen).

During the project we are going to assemble the complete genome and transcriptome of F.candida and O. cincta by using several available assembly platforms (CLC, AbySS). Subsequently, gene annotation and phylogenetic analysis will allow us to further characterize the genome content.

We hypothesize, that occupation of different niches influenced the evolution of stress defense systems. Not only evolution of genes families will be investigated, but also cis-regulatory motives among stress-response gene networks will be analyzed to study plasticity of stress response pathways.
Approach to in silico filtering of primary SNP-call sets in bacterial genome

Nicholas Costa Barroso Lima, Maurício Egídio Cantão, Marisa Fabiana Nicolás

LNCC, Brazil

Abstract: Background: Single Nucleotide Polymorphisms (SNPs) are nucleotide substitutions per site between two sequences. Currently, several studies are focusing on the detection of SNPs in bacterial genomes. To analyze data from the NGS, a bioinformatics approach is mandatory to achieve the SNP-call sets. Thus, a pipeline for detection and filtering of primary SNP-call sets in bacterial genome was developed.

Methods: Scripts in PERL were written to function as filters for the primary SNP-call sets. These filters function in three levels returning a more reliable SNP-call set. SNPs were also classified regarding their location on coding sequence (CDS) or intergenic region. The pipeline was used to detect SNPs among four Klebsiella pneumoniae strains. The reference genome was the clinical isolate Kp13 (Ramos et al., BMC Microbiology, 12:173, 2012) and the comparisons were made against two pathogenic strains (MGH 78578 and NTUH-K2044) and a nitrogen fixing strain (342). SNPs between paired comparisons of these strains were further investigated to assess association with virulence and bacterial fitness.

Results: Three primary SNP-call sets were generated from the comparisons Kp13 vs 342, Kp13 vs MGH 78578 and Kp13 vs NTUH- K2044. Then, from these primary SNP-call sets we obtained nine filtered SNP-call sets. SNPs prevalence is almost ten-fold higher in Kp13 vs 342 than those in the other two comparison analyses. This result may be explained by the different lifestyle of 342 (endophytic plant association). By Poisson Probability Distribution are expected 0-4 SNPs per CDS for Kp13 vs 342 comparison and 0-2 SNPs for Kp13 vs NTUH-K2044 and Kp13 vs MGH 78578 comparisons. Interestingly, CDSs in Kp13 with more SNPs than expected by Poisson Probability are associated with plant growth hormones in comparison with nitrogen fixing strain 342, T4SS and siderophore in comparison with NTUH-K2044 and antimicrobial resistance in comparison with MGH 78578.
P5

The Human Gene Connectome: A Map of Short Cuts for Morbid Allele Discovery

Yuval Itan, Shen-Ying Zhang, Guillaume Vogt, Avinash Abhyankar, Lluis Quintana-Murci, Laurent Abel and Jean-Laurent Casanova

The Rockefeller University, INSERM, Necker Hospital for Sick Children, Pasteur Institute

Abstract: High-throughput genomic and proteomic data obtained by next generation sequencing, microarray or mass spectrometry reveal thousands of gene variants per individual and it is often difficult to decipher which of these variants underlie disease or phenotype in individuals. At the population level however, there can be some level of phenotypic homogeneity, with alterations of specific physiological pathway underlying the pathogenesis of a particular disease. Here we describe the human gene connectome (HGC) as a new approach facilitating the interpretation of abundant genomic and proteomic data, guiding subsequent experimental investigations. We identify the set of shortest plausible biological distances, routes, and degrees of separation between all pairs of human genes, by applying a shortest distance algorithm to the full human gene network. We demonstrate a hypothesis-driven application of the HGC in which we generate a TLR3-specific connectome, which may be useful for the genetic dissection of herpes simplex virus encephalitis of childhood. We also developed the functional genomic alignment (FGA) approach from the HGC. In FGA, the genes are clustered according to their biological proximity (rather than the traditional evolutionary genetic distance), as estimated from the HGC. The HGC and FGA data and computer programs are freely available for non-commercial users, and should facilitate the genome-wide discovery and experimental validation of disease-causing alleles.
P6

Evolutionary age-structured cells dynamics models

Vitalii Akimenko, Yurii Zahorodnii.

Taras Shevchenko National University of Kyiv, Ukraine

Abstract: Age-structured models describe the behaviour of different evolutionary biological systems including the process of cell dividing or birth of living organism. Corresponded researches can be involved in the projects of modelling of growth in physical and biological systems in the cells scale. These models are based on the systems of linear partial differential equations of hyperbolic type – transport equations with special integral boundary conditions. In particular case of linear system for cell density distribution one obtained an analytical solution. This approach allows us to analyse basic lows of behaviour of macroparameters of isolated biological systems and resolve the problems of biological parameters identification and optimal control and get analytical solutions for important applied systems.

P7

A computational model of translation

Dominique Chu and Tobias von der Haar

University of Kent, United Kingdom

Abstract: We present a computational model of translation. This model is genome wide and takes into account all the important steps of translation, including initiation, elongation, ribosome recycling and ribosome-ribosomes interactions on transcripts. In this talk we will first present the software and show how it can be parameterised for any organism. We will then present detailed results obtained for a parameterisation for yeast and also show some experimental results that confirm the predictions of the model.
P8

Modeling the Structure of Human Interferon Gamma C-Termini

Peicho Petkov, Elena Lilkova, Nevena Ilieva, and Leandar Litov

University of Sofia "St. Kliment Ohridski", Sofia

Abstract: Human interferon gamma (hIFN-g) is an important signaling molecule, which plays a key role in the formation and modulation of immune response and in gene activity regulation. The protein represents a non-covalent homodimer of 144 amino acids long monomers, organized in six alpha-helices, which are linked by short unstructured regions, followed by a long unstructured C-terminal domain (the last 21 amino acids). The hIFN-g expresses its activity through the high-affinity interaction with its own species-specific extracellular receptor (hIFNgR). Three regions of the hIFN-g molecule are responsible for receptor binding, one of which is a short sequence in the C-terminal tail (residues 128-131). The hIFN-g C-terminal region is highly positively charged and therefore highly susceptible to proteases. Experimental studies demonstrated that removal of up to 9 residues causes gradual increase in the activity of the protein. However, the complete removal of the flexible C-terminus inactivates the cytokine. Although all studies undoubtedly proved the modulating effect of the unstructured C-terminal region on hIFN-g activity, they failed to explain the molecular mechanism of its action due to the lack of X-ray diffraction data. The extremely controversial conclusions about the function of the hIFN-g C-terminus as well as the lack of a reasonable model explaining its role in the receptor binding prompted us to model its structure by means of molecular dynamics (MD) simulations. This model was then used for investigating the interaction of hIFN-g with hIFNgR1 to reveal the mechanism of this interaction and to shed more light on the role of the C-terminus in hIFN-g -- hIFNgR1 binding.

We performed extensive MD folding simulations (200 ns each), which unambiguously show that the C-termini tend to adopt a compact conformation together with the globular part of the cytokine. The obtained trajectories were subjected to cluster analysis in order to detect the population intensity of the various conformational states. Analysis of the solvation free energy of the centroids of the most populated clusters confirmed their energetic favourability. For validation of the obtained results, the same simulations were performed using different force fields (GROMOS 53a6, AMBER 99SB, and CHARMM 27) with different simulation packages (GROMACS 4.5.4 and NAMD 2.9). Despite the expected minor variations in the details, our conclusions remained unchanged.

Support from Grants DO02-115/15.12.2008, DCVP 02/1/29.12.2009, DIID-02-30/2009 and DRG 02/05/2010 of the Bulgarian National Science Fund is acknowledged. Calculations were performed at the Bulgarian Supercomputing Center and a local Linux cluster at the Atomic Physics Department of the Faculty of Physics of University of Sofia “St. Kliment Ohridski”
An agent-based simulation framework for complex systems

A. Benso, S. Di Carlo, G. Politano, A. Savino, H. Ur Rheman

Politecnico di Torino, Italy

Abstract: In this abstract we present a new approach to the simulation of complex systems as biological interaction networks, chemical reactions, ecosystems, etc... It aims at overcoming previously proposed analytical approaches that, because of several computational challenges, could not handle systems of realistic complexity.

The proposed model is based on a set of agents interacting through a shared environment. Each agent functions independently from the others, and its behavior is driven only by its current status and the “content” of the surrounding environment. The environment is the only “data repository” and does not store the value of variables, but only their presence and concentration.

Each agent performs 3 main functions:
1) it samples the environment at random locations
2) based on the distribution of the sampled data and a proper Transfer Function, it computes the rate at which the output values are generated
3) it writes the output “products” at random locations.

The environment is modeled as a Really Random Access Memory (R2AM). Data is written and sampled at random memory locations. Each memory location represents an atomic sample (a molecule, a chemical compound, a protein, an ion, ...). Presence and concentration of these samples are what constitutes the environment data set.

The environment can be sensitive to external stimuli (e.g., pH, Temperature, ...) and can include topological information to allow its partitioning (e.g. between nucleus and cytoplasm in a cell) and the modeling of sample “movements” within the environment.

The proposed approach is easily scalable in both complexity and computational costs. Each module could implement a very simple object as a single chemical reaction or a very complex process as a gene translation into a protein. At the same time, from the hardware point of view, the complexity of the objects implementing a single agent can range from a single software process to a dedicated computer or hardware platform.
Using Boolean Networks to Model Post-transcriptional Regulation in Gene Regulatory Networks

A. Benso, S. Di Carlo, G. Politano, A. Savino, H. Ur Rehman

Politecnico di Torino, Italy

Abstract: In Gene Regulatory Networks research there is a considerable lack of techniques and tools to understand these networks from a System Biology point of view. The typical biological approach is to reconstruct a particular network from expression patterns, then try to validate its dynamics by simulation, use simulation to analyze its reactions to some perturbations, and finally go back “in vitro” to validate the simulation results. Nevertheless, when the goal is to understand the high-level, general mechanisms that allow these networks to work or to be stable under mild perturbations, this type of approach has shown very strong limitations. In this work we want to better understand the role of miRNA as a stabilizing mechanism in gene regulatory networks. Boolean networks have been recently used to better understand the structural and dynamical properties of regulatory networks. Attractors and ergodic sets have been easily correlated with many of the typical biological cell behaviors (cancer, differentiation, pluripotential, ...). The most widely used model are nevertheless very simple, and work under too strict constraints.

We are defining an enhanced model based on Boolean Networks but also able to take into account post-transcriptional regulation and possibly be extended to other regulatory mechanisms (e.g. ceRNA) that have been already proven crucial in vivo. The final goal is to try to understand if the wide number of miRNA targets constitutes a structural network-stability mechanism used to make the network immune to “regulatory” noise. To achieve this result we evolve the modified Boolean networks for high or low sensitivity to perturbations, and then analyze the resulting networks to understand if specific structural patterns containing miRNA-like post-transcriptional regulatory elements can be correlated with the network stability.
The exploitation of structural and interaction data to predict protein-ligand binding affinity using machine learning

Pedro J Ballester¹, Adrian Schreyer² and Tom L. Blundell²

¹European Bioinformatics Institute, Wellcome Trust Genome Campus, Hinxton - CB10 1SD, U.K. ²Dept. of Biochemistry, University of Cambridge, 80 Tennis Court Rd, Cambridge - CB2 1GA, U.K.

Abstract: Accurately predicting the binding affinities of large sets of diverse molecules against a range of macromolecular targets is an extremely challenging task. The scoring functions that attempt such computational prediction exploiting structural data are essential for analysing the outputs of molecular docking, which is in turn an important tool for drug discovery, chemical biology and structural biology. Conventional scoring functions assume a predetermined theory-inspired functional form for the relationship between the variables that characterise the X-ray structure of the complex and its predicted binding affinity. The inherent problem of this approach is in the difficulty of explicitly modelling the various contributions of intermolecular interactions to binding affinity.

A new family of scoring functions able to effectively exploit much larger amounts of experimental data and circumvent the need for an explicit functional form has recently been introduced¹. These machine learning regression models have been shown to outperform a wide range of state-of-the-art scoring functions in a publicly available benchmark². Here we will investigate the impact of the way the protein-ligand complex is encoded on the predictive power of the resulting scoring function. This systematic battery of numerical experiments has led to an even more accurate scoring function. An additional contribution of this study is in the critical assessment of modelling assumptions. In the light of our performance results and considering the uncertainty introduced by the static nature of crystal structures, we will discuss the role of interatomic distance cutoffs and binning as well as protonation states for the binding affinity prediction of a protein-ligand complex.
P12

CASoX - a software tool to visualize and analyze cavities and clefts in protein structures

Georg Steinkellner, Gustav Oberdorfer and Karl Gruber

ACIB GmbH, Graz, Austria; Institute of Molecular Biosciences, University of Graz, Austria

Abstract: We have developed a plug-in for the visualization tool PyMOL for a convenient and flexible representation of cavity information. This will lead to a better and enhanced analysis of cavities in distinct cases and will also give new insights into specific structural behaviors like water channels or exit tunnels. Cavity and cleft information can be very useful to understand biologically or industrially important protein behaviors. The shape of the cavity as well as the size, solvent accessibility and physico-chemical properties are often relevant for binding and conversion of enzyme specific substrates. A convenient calculation, visualization and interaction tool for cavities and clefts will ultimately help to analyze, interpret and present cavity information more easily. Most cavity programs output the best proposed cavities for different prediction purposes which are defined by internal program settings. This makes it more likely to miss some cavity features or to interpret cavity or cleft information differently. With CASoX you can show all calculated cavities and select for yourself which cavity or cavities are important merely by looking to the overall shape and location of the cavities or by playing with the standard settings. Therefore also active sites which are not in a closed cavity can be viewed and analysed. It is easier to identify water or substrate tunnels which are not directly visible because e.g. an amino acid may block the entry (or exit). It is also possible to calculate and map hydrophobicity and accessibility values onto the cavity points. (Figure 1) By utilizing the capabilities of PyMOL every calculated cavity can be viewed and every cavity point is accessible for selection. Presets are provided to use the tool very quickly without going deeper into the advanced features and settings, providing a fast cavity calculation tool to be used as a standard analysis instrument.
P13

Cupin E8WYN5 binding pocket analysis and visualization via CASoX plugin

Andrzej LYSKOWSKI, Georg STEINKELLNER, Kerstin STEINER, Helmut SCHWAB and Karl GRUBER

ACIB GmbH, Graz, Austria

Abstract: The recombinant protein AciX9_0562 from Acidobacterium sp. MP5ACTX9 (UniProt ID E8WYN5) has recently been crystallized and preliminary structure has been calculated. A structure of 131 amino-acid protein of uncharacterized function with a molecular weight of 14.25 kDa reveled a large cavity with a putative active site located around a metal ion bound at the bottom. Here we present a novel cupin E8WYN5 structure and a visualization of binding pockets and the preliminary analysis via a CASoX PyMOL plugin. The graphical interpretation of the cavity volume as well as its biophysical properties are presented in easy to interpret graphical form. The standard setting of the CASoX plugin allow for the rapid visualization of multiple properties including but not limited to cavity proportions and its hydrophobicity. Ability to rapidly visualize the properties is essential to interpret and analyze specific activity of the enzyme particularly in case when the exact native activity is unknown.
P14

Prediction of intra-communication pathways within the second PDZ domain of PTP1e

Elisa Cilia (1), Geerten W. Vuister (2) and Tom Lenaerts (1,3)

(1) MLG, Département d’Informatique, Université Libre de Bruxelles, Boulevard du Triomphe CP212, 1050 Brussels, Belgium (2) Department of Biochemistry, University of Leicester, Henry Wellcome Building, Lancaster Road, Leicester LE1 9HN, United Kingdom (3) AI-lab, Vakgroep Computerwetenschappen, Vrije Universiteit Brussel, Pleinlaan 2, 1050 Brussels, Belgium

Abstract: Experimental NMR relaxation studies have shown that peptide binding induces dynamical changes at the side-chain level throughout the second PDZ domain of PTP1e, identifying as such the collection of residues involved in long-range communication. Even though different computational approaches have identified subsets of residues that were qualitatively comparable, no quantitative analysis of the accuracy of these predictions was thus far determined. We show that our information theoretical method produces quantitatively better results with respect to the experimental data than some of these earlier methods. Moreover, we provide a global network perspective on the effect experienced by the different residues involved in the process. We show furthermore that these predictions are consistent within both the human and mouse variants of this domain. Together, these results improve the understanding of intra-protein communication and allostery in PDZ domains, underlining at the same time the necessity of producing similar data sets for further validation of this kind of methods.
P15

The Effects on Sensitivity of the Structure-Activity Models

Sorana D. BOLBOACĂ¹ and Lorentz JÂNTSCHI²

¹"Iuliu Hațieganu" University of Medicine and Pharmacy Cluj-Napoca, Department of Medical Informatics and Biostatistics, 6 Louis Pasteur, 400349 Cluj-Napoca, Cluj, Romania. E-mail: sbolboaca@umcluj.ro; ²Technical University of Cluj-Napoca, Department of Physics & Chemistry, 103-105 Muncii Bdv., 400641 Cluj-Napoca, Romania. E-mail: lorentz.jantschi@gmail.com

Abstract: Aim: Model sensitivity analysis was conducted to assess the effects on statistical characteristics of linear regression model obtained to link the toxicity with compounds hydrophobicity and energy of the lowest unoccupied molecular orbital. Material and Method: A sample of 250 phenolic compounds with measured toxicity on Tetrahymena pyriformis, calculated octanol/water partition coefficient (logP) and LUMO (energy of the lowest unoccupied molecular orbital taken from [Cronin et al., Chemosphere 2002;49:1201-1221]) were taken from a previous published manuscript [Zhao et al., Chemosphere 2009;75(7):866-871]. Standardized residuals (values higher than 2.5 [Tropsha et al., QSAR Comb Sci 2003;22(1):69-77] and 3 [Gramatica, QSAR Comb Sci 2007;26(5):694-701]), hat-matrix leverage (h_i>2*(k+1)/n, where h_i = leverage for i-th compound, k = number of independent variables in the regression model and n = sample size), and Cook's distance (D_i>0.85 (k=2) [Mc Donald, Res Lett Inf Math Sci 2002;3:127-128]; D_i>1 [Cook and Weisber, Residuals and influence in regression, 1982] and D_i>4/n [Bollen and Jackman, Regression diagnostics: An expository treatment of outliers and influential cases, 1990:257-291]). The compounds identified with values of residual, hat-matrix leverage and Cook's distance higher than thresholds were remove and the model was rebuilt. The removal was performing whenever an improvement in correlation coefficient was obtained. Several statistical criteria were used to validate and to compare models. Steiger's test [Steiger, Psychol Bull 1980;87:245-251] at a significance level of 5% was applied to test if correlation coefficients were statistically different between models. Results: Basic statistical characteristics of the full model (n=250) were as follow: R²=0.5643, R²_adj=0.5608, RMS=0.5491, F-value=160, p-value=2.77·10^-45; all coefficients in the model statistically significant - p-value<0.001; R²_cv_loo= 0.5513. Two compounds were with standardized residuals higher than 3, but since these compounds were also identified by 2.5 threshold no models were constructed on this criterion. The removal of compounds with h_i higher than threshold did not led to any improvement of the model characteristics and was not further investigated. Two models were further analyzed to compare theirs abilities in estimation and prediction: one model with 219 compounds obtained after removal of compounds with r>3 (R²=0.8530, R²_adj=0.8517, RMS=0.3121, F-value=627, p-value=1.51·10^-30; all coefficients in the model statistically significant - p-value<0.00001; R²_cv_loo=0.8482) and one model with 179 compounds obtained after removal of compounds with D_i>4/n (R²=0.9099, R²_adj=0.9089, RMS=0.2352, F-value=889, p-value=9.94·10^-33; all coefficients in the model statistically significant - p-value<0.00001;
$R^2_{\text{cv-loo}}=0.9068)$. D-model proved to be most reliable model compared to full-model and $r_i$-model according to applied validation criteria (100% accomplished). The $r_i$-model proved a significant higher correlation coefficient compared to full model ($Z=2.564, p=0.0052$). D-model proved having the higher correlation coefficient compared to both full-model and $r_i$-model ($Z>6.841, p<3.96 \cdot 10^{-12}$). **Conclusion:** The Cook’s distance approach proved the method able to identify those compounds with significant influence on the QSAR model, providing a reliable and valid model compared to both full-model and model obtained by applying the residual approach. **Keywords:** quantitative structure-activity relationships (QSAR); residuals ($r_i$); hat-matrix ($h_i$); Cook’s distance ($D_i$)
P16

Natural compound extracts diversity among genus

Lorentz JÄNTSCHI¹ and Sorana D. BOLBOACĂ²

¹ University of Agricultural Science and Veterinary Medicine Cluj-Napoca, 3-5 Calea Mănăștur, 400372 Cluj-Napoca, Romania. E-mail: lorentz.jantschi@gmail.com; ²“Iuliu Haţieganu" University of Medicine and Pharmacy Cluj-Napoca, Department of Medical Informatics and Biostatistics, 6 Louis Pasteur, 400349 Cluj-Napoca, Cluj, Romania. E-mail: sbolboaca@umfcluj.ro

Abstract: Aim: The aim of our research was to evaluate the intrinsic diversity of genus based on chemical composition of natural extracts of its constituent species. Material and Method: Five genuses were included in analysis: Cacalia, Dracocephalum, Jatropha, Saussurea and Senecio. The number of known active compounds and the species where the compounds appear were taken from previous published materials. The Bootstrap method was used to estimated based on independent observation (compounds obtained from plant extracts) the distributions of a series of statistics (such as mean number of chemical compounds, entropy (Shannon and max-entropy) and diversity (Simpson's diversity index) for genus, number of plants and/or chemical compound) in order to illustrate what happen in population. The simulation study was conducted by 10,000 times for each number of plants varying from 1 to 50. Results: The minimum number of active compounds (4.10) was observed for Senecio when just one plant was included in analysis. As expected the maximum number of compounds (243.35 - Dracocephalum) was observed when 50 plants were included in analysis. The patterns of richness in natural compounds according to genus proved similar with the pattern of Simpson's diversity index. Shannon's entropy had values from 1.1259 (Senecio - n=1) to 5.3397 (Dracocephalum - n=50). Without any exception, the Shannon's entropy slightly increased with sample size but the increased proved not uniform when the differences were analyzed. The entropy of most frequent compound (max-entropy) varied from 1.1259 (Senecio - n=1) to 4.0842 (Dracocephalum - n=50). The max-entropy systematically increased with sample size just for Senecio. The rule of increase of max-entropy with sample size was broken by 1 (Saussurea), 2 (Jatropha), 3 (Cacalia) and respectively 9 times (Dracocephalum). Conclusion: The obtained results revealed that both in terms of richness in active compounds and weight of most frequent extracted compound Dracocephalum genus is most suitable for propagation being followed by Saussurea, Cacalia and Jatropha prove the lowest diversity in terms of active chemical compounds. Keywords: biodiversity; bootstrap; rarefaction; compounds richness
P17

Metadynamics simulations of FRET efficiency in dye-labelled polyproline system

Katarzyna Walczewska-Szewc, Ben Corry

The University of Gdańsk, Poland; The Australian National University, Australia

Abstract: Forster Resonance Energy Transfer (FRET) is the radiationless transfer of energy occurring between excited molecule (donor) and ground-state molecule (acceptor) via long range dipole-dipole interaction. It is commonly used to study the structure and dynamics of proteins by attaching chromophores to known sites and measuring the efficiency of transfer between them. However, the measured transfer is dependent on both the separation and orientation of the participating molecules and it is difficult to experimentally isolate these effects.

We are using molecular dynamics (MD) simulations to replicate the conditions of experimental measurements in order to enable the structural basis of the measured transfer efficiency to be determined. In order to do this, enhanced sampling methods such as metadynamics will need to be employed to see all the possible conformational states that can yield given FRET efficiency values. First, to test the accuracy of the method, it is applied to the simple case of two freely rotating chromophores at the fixed distance, where results can be compared to analytical data. At the next step, metadynamics is used in a well characterized system of dye-labelled polyproline. The results of numerical simulations are compared with the experimental ones.
P18

Bioinformatic analysis of complete genome sequence of Bacillus atrophaeus UCMB-5137 used as a bio-pesticide

Oleg N. Reva, Wai Y. Chan, Candice N. Ryana, Svitlana V. Lapa, Lilija V. Avdeeva, Rainer Borriss

Department of Biochemistry, Bioinformatics and Computational Biology Unit, University of Pretoria, Gauteng, South Africa; Department of Microbiology and Plant Pathology, University of Pretoria, Gauteng, South Africa; Rhodes University Bioinformatics (RUBi), Department of Biochemistry, Microbiology and Biotechnology, Rhodes University, Grahamstown, South Africa; Department of Antibiotics, Institute of Microbiology and Virology, Kiev, Ukraine; ABiTEP CmbH, Berlin, Germany;

Abstract: Using NGS technologies for complete bacterial genome sequencing have become a common practice. With the reduced cost in NGS, genome sequencing is financially affordable even for small laboratories. However, for many potential users it remains unclear how to proceed with this huge amount of genetic data released by NGS and how this information may be linked to real practical questions. Genetic barcoding of microorganisms was the first area where new sequencing technologies met the requirements of the applied microbiology. DNA barcoding in microbiology is a genome comparative approach aiming to identify the difference between bacterial strains, which are hardly distinguishable by any traditional methods. The following questions may be addressed: i) is this strain unique and if so, what makes this strain so effective in its biotechnological application; ii) how this strain may be traced down in the environment; iii) are there any genetic markers of its extraordinary biological activity? In this work we present new bioinformatics tools and approaches developed to address these questions. The computational tools were benchmarked on a newly sequenced genome of Bacillus atrophaeus UCMB-5137 used in biotechnology as a bio-pesticide for plant and crop protection against phytopathogens. The strain was compared against available sequences of B. atrophaeus and B. amyloliquefaciens genomes. The latter once are well known endophytic plant growth promoting bacteria, but this activity was not reported for other B. atrophaeus. Genome sequence analysis showed that the unique biological activity of UCMB-5137 was related to several horizontally transferred genes, which were found also in other plant associated bacteria. Inhabiting of plant rhizosphere had promoted further adaptation changes in many genes of this bacteria that led to a phylogenetic diversion of this strain from other B. atrophaeus isolates.
P19

Elucidating the nature and origin of the biopesticidal strain B. atropheus UCMB-5137

Rian Pierneef, Prof. Oleg Reva, Prof. Rainer Borriss

University of Pretoria, South Africa

Abstract: The use of biotechnology in the production of consumables is of critical importance in a resource dependent environment. The use of bacterial strains in biotechnology requires the identification and investigation of horizontally transferred genes in order to understand their origin and activity. The SeqWord Gene Islands Sniffer (SWGIS) is used to predict horizontally transferred genes and Genomic Islands (GIs) through pattern distribution and optimized parametric measurements. A database of considerable size has been developed to house GIs predicted by SWGIS with innovative tools to identify homologous GIs and their possible origin. These tools have been used in combination to study and elucidate the genome of the naturally pesticidal strain, Bacillus atropheus UCMB-5137. The specificity of the strain is clearly related to the specificity of the strain's GIs. The majority of GIs of B. atropheus UCMB-5137 are unique in comparison to the closest relative strain B. atropheus 1942. The origin of these GIs was found to be of plasmids harboured by distant rhizosphere bacteria.

The combination of GI prediction tools and a large database clearly illustrates the power and application of bioinformatics in the field of biotechnology. The specificity of GIs and the inferred specificity on the host are of great importance in the understanding and application of bacteria with regard to biotechnology.
Analysis of predicted thermostable proteins for potential biotechnology application in Thermus bacteria

Benjamin Kumwenda, Derek Litthauer and Oleg Reva

University of Pretoria, Free State University, South Africa

Abstract: Enzymes from thermophilic bacteria exhibit higher activity and stability than mesophilic or synthetic enzymes currently being used in industry for production of food, clothing, laundry, shoe and paper. Thermostable enzymes do not denature in high temperatures and are also tolerant to extreme acidic or alkaline conditions. This cancels the need for expensive cooling machines in industry, enables longer and high reaction rates due to high substrate solubility resulting into more productivity and eventual profits. The application of next generation sequencing techniques in metagenomics has allowed the isolation of new genes that encode for industrial potential thermostable enzymes. However, the prediction of protein thermostability from wet laboratory experiments is not only time consuming but also costly due to expensive equipment, reagents and the expertise required. In silico prediction of thermostability also has its challenges because factors on which prediction methods are based are rather contradictory as they are either organism or protein specific. In this work, we have used Thermus genomes and proposed a new approach based on UNAFold algorithm for RNA secondary structure prediction to determine thermostable enzymes. Analysis show a consistent difference in thermostability as determined by minimum folding energy (kal/mol) between mRNAs encoding for proteins and enzymes from thermophilic and mesophilic bacteria. A strong correlation between mRNA adaptation and positive selection of amino acid substitutions in encoded proteins has been reported. The new approach and a Python program are available at www.bi.up.ac.za/SeqWord/gct/index.html for genome comparison and prediction of thermostable proteins and genes.
P21

Analysis of microRNA array expression data: evidence for false-positive signals

Mitiushkina N.V.\(^{(1)}\), Iyevleva A.G.\(^{(1,3)}\), Kuligina E.Sh.\(^{(1)}\), Togo A.V.1\(^{(1)}\), Miki Y.\(^{(2)}\), Imyanitov E.N.\(^{(1,3)}\)

\(^{(1)}\) Laboratory of Molecular Oncology, N.N. Petrov Institute of Oncology, St.-Petersburg, Russia; \(^{(2)}\) Department of Genetic Diagnosis, Cancer Institute, Japanese Foundation for Cancer Research, Tokyo, Japan; \(^{(3)}\) Department of Medical Genetics, St.-Petersburg Pediatric Medical Academy, St.-Petersburg, Russia

Abstract: MicroRNAs became recently known as important biomolecules taking part in many vital processes such as cell growth, differentiation or malignant transformation. MicroRNA profiling via oligonucleotide arrays is a widely used technique to study microRNAs. We would like to draw attention to a specific problem which may arise during statistical analysis of microRNA expression data. The problem is that certain microRNAs, which are expected to have low or absent expression, nevertheless produce strong signal in particular samples.

The data were collected for the study of bilateral and unilateral breast tumour features. Agilent microRNA arrays were used to obtain microRNA profiles of 84 formalin fixed paraffin embedded (FFPE) tumour tissue samples. These arrays contained probes for more than 1,000 human microRNA species, as well as probes for human viral microRNAs. Hierarchical clustering of microRNAs using Pearson’s correlation coefficient as a similarity measure revealed two main clusters. Several evidences made us believe that microRNAs belonging to one of these clusters (“cluster 1”) are in fact not expressed at all in our samples. First of all, it contained 25 of all 26 viral microRNAs found to be expressed in our samples, including those attributed to HIV, HBV and other viruses very unlikely to be really present in mammary tissue. Second, it mostly consisted of microRNAs with large name numbers (p<2.2\(\times\)10-16 when compared to microRNAs from other cluster, “cluster 2”, Mann-Whitney U test). Large number in a name of a microRNA indicates its more recent discovery and usually is associated with its low level of expression. On the other hand, microRNA name numbers and the percentage of viral microRNAs in this cluster were not different from those of microRNAs with no detectable expression. This leads to suggestion that “cluster 1” microRNAs present just a random sample of non-expressed microRNAs. We also analysed nucleotide sequences of microRNAs forming different clusters. It appeared that sequences of microRNAs from “cluster 1” were extremely rich in guanine (p<2.2\(\times\)10-16 when compared to “cluster 2” or non-expressed microRNAs sequences, Mann-Whitney U test). Moreover, the level of expression of these microRNAs showed positive correlation with their guanine content (median Spearman’s correlation coefficient 0.389, p=6.51\(\times\)10-7). No correlation of “cluster 2” microRNAs expression with their nucleotide composition could be found.
It is still unknown, however, why probes for guanine-rich microRNAs tend to produce strong signal on oligonucleotide arrays even in the apparent absence of their targets and why this effect became more pronounced in some samples. Results of a recent study by Borgan E. et al. [Mol Oncol, 2011, Vol. 5(6)] look rather intriguing in this context. With the same Agilent microarray technique they showed that delayed freezing of tumour tissues induce expression of a number of microRNAs. Almost all of these microRNAs in our study fell into “cluster 1”, the rest of them were not found to be expressed. We do not have information about time to formalin fixation of our tumour samples, but it’s highly possible that this time varied significantly and may affect subsequent results. The other interesting observation made in our study is that the mean expression of “cluster 1” microRNAs was significantly higher in estrogen-negative tumors than in estrogen-positive tumors (p=0.0007, Student’s t test). Further investigation is needed to find out if the increase in guanine-rich sequences in some tissues is a real biological phenomenon or is this all a result of some technical complication. It is important though to be aware of the existence of such a problem. Our experience showed that it can interfere with normalization and subsequent analysis of the microRNA expression data.
P22

PIK3CA mutation screening in Bulgarian breast cancer patients

Rumyana Dodova1, Stefan Bozhanov2, Ilia Nikolov2, Zinaida Garanina3, Atanaska Mitkova1, Radka Kaneva1, Vanio Mitev1, Elena Georgieva2

1 Molecular Medicine Center and Department of Medical Chemistry and Biochemistry, Medical University – Sofia; 2 Institute of Plant Physiology and Genetics, Akad. “G. Bonchev”, Bl. 21, 1113 Sofia, Bulgaria, Department of Molecular Genetics, Bulgarian Academy of Sciences; 3 National Oncology Hospital, Bulgaria - Sofia

Abstract: Studies of different populations, demonstrate that somatic mutations in PIK3CA proto-oncogene account for sporadic breast, ovarian, colorectal, gastric and hepatocellular cancers. Our previous studies in sporadic breast cancer patients of Bulgarian origin demonstrate high frequency of PIK3CA mutations in exon 9 and 20. We performed screening of all exons of PIK3CA by direct Sanger sequencing in 38 genomic DNAs isolated from fresh frozen breast tumours for mutation in PIK3CA (6 familial and 32 sporadic cases) and 78 healthy control women. The results were analyzed by SeqScape software and Pearson chi square test. Our results demonstrate the presence of the 3 hot spot mutations E542K, E545K, H1047R, with pathological effect. The combined frequency of the mutations E542K, E545K and H1047R in breast cancer patients is 13% (one patient with E542K, two with H1047R and two with E545K). This is lower than then the previously reported results in sporadic breast cancer of Bulgarian patients (31,25%), most likely due to the small number of the patients investigated. The identified mutations do not show correlation with the age, grade and other clinical characteristics of the tumours. The mutations in exon 20 are found in ER/PR positive and Her2 negative tumours. A common polymorphism (A1173G) I391M in exon 6, coding for the C2 domain of the protein has been found. The frequency of the rare G allele is 15,8 % in patients and 7,1 % in healthy controls. The presence of one G allele in I391M is associated with increased risk for breast cancer (p=0.05, OR= 2.47). No genotypes show significant association, most likely due to the small sample size. To our knowledge, there are no available data in the literature about the association of this polymorphism with breast cancer. Further analysis in larger population sample is required to confirm or exclude our preliminary findings.

Acknowledgements: We declare that the study detailed in the abstract is funded by Bulgarian Science Fund, Contract #: DO-02-310/08
The impact of saponins from Tribulus terrestris on gene expression in breast cancer cells

Stefan Bozhanov*, Teodora Goranova#, Valentin Lozanov&, Radka Kaneva#, Ilia Nikolov*, Vanio Mitev#, Gerardo Lopez-Rodas^ and Elena Georgieva*

*Institute of Plant Physiology and Genetics, Akad. “G. Bonchev”, Bl. 21, 1113 Sofia, Bulgaria; #Molecular Medicine Center, Medical University – Sofia, 1431 Sofia, Bulgaria; &Department of Medical Chemistry and Biochemistry, Medical University – Sofia, 1431 Sofia, Bulgaria; ^Department of Biochemistry and Molecular Biology, University of Valencia, 46100 Burjassot, Valencia, Spain

Abstract: Saponins are a group of glycosides present in various higher plants. Recently numerous studies are undertaken on them since they are found to possess anti-tumour properties. Saponins from Tribulus terrestris (TT) are reported to have anti-cancer effect on various cancer types. In the present study we evaluated the impact of saponins from TT on gene expression of 30 up- and down-regulated genes in breast cancer cell line MCF-7. The selected group included genes from various pathways responsible for cell migration and metastasis formation, apoptosis, DNA repair, DNA demethylation, regulation of the cell cycle, transcription and angiogenesis. Expression of the genes was quantified in MCF-7 cells before and 3, 7, 24, 48 and 72 hours after the treatment with saponin extract from TT using real-time PCR with SYBR Green. The expression level of each gene was normalized to the respective level of beta-actin in each sample.

No significant change in the expression of most of the genes before and after treatment was found. However, dynamics in the mRNA level of CXCR4 was observed. The chemokine receptor CXCR4 is involved in the migration, invasiveness, metastasis and proliferation of breast cancer cells. Seven hours after the application of saponins from TT there was a significant decrease in CXCR4 expression and it was still under-expressed at 24 hours; afterwards it gradually increased reaching the mRNA level before treatment.

As saponins vary in structure, possibly they could exert their anti-tumour properties affecting variety of anti-tumour pathways. Our results suggest that saponins from TT affect CXCR4 pathway and thus they are likely to inhibit cancer cell metastasizing.

Acknowledgements:
We declare that the study detailed in the abstract is funded by Bulgarian Science Fund, Contract #: DO-02-310/08
Abstract
Sugar beet [Beta vulgaris (L.) subsp. vulgaris] is an important crop for sugar production, providing approximately a third of all sugar consumed in the world. In this study, we evaluated the distribution and mapping of potential SNPs and their role as markers for fingerprinting of the sugar beet germplasm by means of a high-throughput marker array technology (TaqMan open array, Life Technologies, Carlsbad, CA). We mapped a potential set of 838 SNP pairs and it was found that only 225 SNP pairs mapped perfectly with one bp alteration. Among the 225 mapped SNPs, we observed that the transitions ([A/G]+[C/T] represented the most abundant changes (115; 58.67%), which were followed by transversions ([A/C]+[G/T], 81;41.33%). We further verified a selected set of 192 SNPs using the TaqMan open array and we found that the selected SNPs were able to efficiently resolve a set of sugar beet varieties. Also, the observed phylogenetic distribution were congruent to previous reports. Therefore, we postulated a wide repertoire of SNPs which could serve as a potential source for evaluating the sugar beet germplasm.
Abstract: The application of Next Generation Sequencing to plant research made large-scale analyses of plants without sequenced genomes feasible. However, the complexity of plant genomes with a history of repeated whole genome duplications and reductions, tandem duplications and propagation of repetitive elements complicates sequencing approaches towards plant genomes. Plant transcriptomes in contrast represent a valuable resource, in particular for non-model species without a sequenced genome. To date, two different strategies address the challenge of assembling short sequencing fragments de novo to full-length sequences (contigs), overlap consensus based assemblers for few, long reads and de-Bruijn graph based assemblers for many, short reads. Initially all the tools for de novo assembly were designed for - and tested on - assembling genomes. Due to the high dynamic range of transcript abundance, about five orders of magnitude, transcriptome assemblies are still far from resembling all transcripts of the actual plant. Modifications of the software towards de novo assembly of transcriptomes were developed, but not thoroughly evaluated. In addition to reliable contigs current de novo assemblers produce a variety of incorrect contigs including those that are too short, too long and contigs that are assembled from two distinct transcripts.

We have developed tools to assess the quality of transcriptome assemblies. A prototype pipeline to detect chimeric contigs has been developed and is available as open source. We show that assessing the quality of a plant transcriptome in a comparable, comprehensive way is easily achieved and should be included in every de novo transcriptome paper. Furthermore, we discuss the implications on development of future transcriptome assembly and unsupervised assessment software.
Comparative genome analyses of pathogenic treponemal species

Pavlína Krečmerová, Helena Pětrošová and David Šmajs

Department of Biology, Faculty of Medicine, Masaryk University, Kamenice 5, Building A6, 625 00 Brno

Abstract: Introduction: Treponema pallidum ssp. pallidum (TPA), the causative agent of syphilis, and Treponema pallidum ssp. pertenue (TPE), the causative agent of yaws, are closely related spirochetes causing diseases with distinct clinical manifestations. While syphilis is a worldwide venereal and congenital disease, yaws is a tropical disease transmitted by direct skin contact. Closely related Treponema paraluiscuniculi (TPC) is the causative agent of rabbit venereal spirochetosis and is not infectious to humans. Differences on a genomic level are minimal between TPA and TPE strains (0.2%) and also between TPA strains and TPC (less than 2%). However, genetic basis and evolution of all treponemal diseases remain unknown. Nowadays, a number of finished whole genome sequences of treponemal species enables comparative genome studies. Here we present results of comparative evolutionary analyses of five TPA strains (Chicago, Dallas Nichols, Mexico A and SS14), three TPE strains (CDC-2, Samoa D and Gauthier) and TPC strain Cuniculi A.

Methods: All genome sequences were taken from the GenBank database under following accession numbers: TPA strains: Chicago CP001752.1, Dallas CP003115.1, Nichols AE000520.1, Mexico A CP003064.1, SS14 CP000805.1; TPE strains: CDC-2 CP002375.1, Samoa D CP002374.1, Gauthier CP002376.1; TPC strain Cuniculi A CP002103.1. All annotated genes were aligned as pairs using global alignment algorithm. All possible combinations of gene alignments were tested on selection by Z-test on selection using modified Nei-Gojobori method, Jukes-Cantor model with complete deletion as gap treatment and transition/transversion ratio set to 0.85. We evaluated type of selection for each annotated gene within TPA and TPE groups and also between TPA and TPE strains, TPA strains and TPC and TPE strains and TPC. Because of high sequence diversity and unidentified nucleotides in consensus sequences in majority of TP genomes, tprK gene was excluded from our calculations. To obtain significant results, a threshold of 5 nucleotide changes per gene was set for positive selection and 2 nucleotide changes for negative one. Results were considered significant at the 5% level.

Results: There are 26 genes under positive selection in at least one comparison of TP strains. 16 of them have predicted biological function: 6 are involved in bacterial virulence, 4 in cell processes and cell structure, 3 in general metabolism, 1 in transcription, 1 in transport and 1 in regulation; 10 encode hypothetical proteins. Negative selection influences 314 genes in at least one comparison of TP strains, and operates most strongly between TPA strains and TPC (278 genes) and also between TPE strains and TPC (251 genes). Interestingly, no genes under negative selection were found within the group of TPE strains, indicating that TPE group is most...
conservative one. Besides genes with unknown function (98 genes), most commonly negatively selected genes are involved in general metabolism (53 genes), in transport (40 genes) and in translation (33 genes). There are only 3 genes which are under positive selection in one comparison of TP strains and at the same time under negative selection in the other – TP0136, TP0344 and TP0859. Two of them have predicted biological function: TP0136 is putative membrane protein involved in virulence and TP0344 encodes transcription-repair coupling factor TrcF. Third one, TP0859, encodes hypothetical protein.

Conclusion: Results from this study will be used for further phylogenetic analysis. Also, genes identified under positive selection represent candidate genes for research of treponemal virulence factors.
P27

Computational identification and analysis of pupylation sites

Chun-Wei Tung

School of Pharmacy, Kaohsiung Medical University, Kaohsiung 807, Taiwan

Abstract: The recently identified prokaryotic ubiquitin-like protein (Pup) is an essential protein modifier for regulating protein functions. Similar to ubiquitin, the well-studied function of Pup is to tag proteins for proteasomal degradation. The identification of pupylation sites can provide insights into the substrate specificity and regulated pathways and functions of pupylation. Compared to expensive and labor-intensive experimental approaches such as tandem mass, this study aims to develop computational methods for the identification and analysis of pupylation sites. To develop a computational model for identifying pupylation sites, a PupDB database was firstly constructed to serve as a comprehensive collection of pupylated proteins and pupylation sites (Tung, 2012). Based on the pupylation sites of PupDB, an efficient support vector machine classifier was developed by using a modified composition of k-spaced amino acid pairs. The developed classifier with area under the receiver operating characteristic curve of 0.83 is better than the only available classifier of GPS-PUP (0.749) (Liu, et al., 2011) using 10-fold cross-validation. The analysis of statistically significant features of k-spaced amino acid pairs shows that pupylation sites near terminals are especially important for discriminating pupylation sites. To better understand functions of pupylation, the developed classifier was applied to identify pupylation sites of proteins, and the involved functions and pathways were analyzed using gene ontology. The results showed that pupylation regulates broad functions other than proteasomal degradation.

References
miR414-MEDIATED TRANSLATION REGULATION OF MYB GENE FAMILY

Assyl Bari, Anatoliy Ivashchenko

Al-Farabi Kazakh National University, Kazakhstan

Abstract: microRNAs (miRNAs) are key regulators of important plant growth and development mechanisms at the post-transcriptional level. In order to characterize miR414 which has a wide range of potential targets, features of miR414 interaction with MYB transcription factor family have been investigated. Our approach consisted of in silico analysis using online databases (Genbank, miRBase, KEGG) and miRNA-target prediction program RNAHybrid 2.1. We searched miR414 targets among 131 members of MYB family. AtMYB4R1, AtMYB12, AtMYB60, AtMYB70, AtMYB77, AtMYB84, AtMYB98 and AtMYB110 have been revealed as miR414 targets in Arabidopsis thaliana. These genes have been found to be involved in the control of processes including primary and secondary metabolism, cell fate and identity, developmental processes and responses to biotic and abiotic stresses. All interaction sites are located in protein-coding sequence (CDS) of mRNA. Also potential miR414-regulated targets among orthologs of MYB genes in Arabidopsis lyrata, Glycine max, Medicago truncatula, Sorghum bicolor, Vitis vinifera have been studied. In nucleotide level miR414 binding sites are characterized by GAN-rich trinucleotide repeat sequences. We further investigated the targets in the protein level and found that the region targeted by the miRNA at the transcriptional level consists of an Asp-rich motif. However, in some cases frameshift and, respectively, replacement of aspartic acid has been observed. This study suggests that interaction of miRNA with protein-coding region of many genes appeared long time ago and has been preserved in evolution process. Conservation of miRNA and binding sites sequences across species provides powerful tool for identification of novel miRNA targets based on homology.
P29

Parameter identification of urea cycle model using evolutionary algorithms

Anastasia Slustikova Lebedik, Ivan Zelinka

Tomas Bata University in Zlin, Czech Republic

Abstract: Background: For many years, computational tools have been applied widely to the study biological systems. In recent years, mathematical modeling and computer simulation of cellular processes has become one of the most developing topics in current bioscience. Due to advances in molecular biology and computer technology, it has become possible to reach valuable insight into biological processes at genes, metabolites, proteins or cell levels. One of the central roles in biological processes plays metabolism. Hence, the modeling of metabolic systems is a key to a complete understanding of normal and pathogenic processes in the human body. Many metabolic system models have been published in recent years. Different methods have been applied to capture the precise models. However, one of the principal questions in modeling of metabolic systems is the model parameters identification. The main source of the kinetic parameters is published literature. Nevertheless, most of the kinetic parameters are not available. Due to the large number of reactions, the non-linear interactions between different metabolites and enzymes, the parameters identification problem is highly complex.

Results: We investigate application of evolutionary algorithms for parameters identification problem. Three evolutionary computation techniques, called Differential Evolution (DE), genetic algorithm (GA) and Self-Organizing Migrating Algorithm (SOMA), are applied to well-studied metabolic system, the urea cycle of the mammalian hepatocyte. The model includes four enzyme reaction schemes: arginase, ornithine carbamoyl transferase, argininosuccinate lyase, and argininosuccinate synthetase. All experiments are conducted using Mathematica 8. The efficacy of particular method is judged based on the minimal, maximal and average values of the parameters in comparison with true values of the parameters from published sources. The cost function value is also used as a quality measure of algorithms performance.

Conclusions: All three optimization techniques yield meaningful results. Each algorithm is capable of finding good local optima. Notably, SOMA especially provides good performance. Considering computational time, the most time-consuming calculations are observed in SOMA case. In contrast to SOMA, DE and GA experiments take less time. The algorithms provide an effective approach in parameters identification problem. Taking into account that all algorithms are performed well, and the fastest is DE and GA, it is reasonable to apply DE in experiments with limited computational time. Nevertheless, the performance of SOMA is a significant point in our experiments.
First glance at the genome of the club root pathogen Plasmodiophora brassicae

Johan Fogelqvist, Arne Schwelm and Christina Dixelius

Swedish University of Agricultural Sciences, Sweden

Abstract: Plasmodiophora brassicae is the casual agent of the clubroot disease in Brassicaceae, one of the most damaging diseases within this plant family. Despite its agricultural importance, the biology of P. brassicae remains poorly understood. Due to its obligate biotrophic nature, P. brassicae remains impossible to grow in axenic culture and the typical experimental systems for working with P. brassicae are comparatively unsophisticated. Molecular studies are especially challenging with only approximately 100 known genes. We recently succeeded in obtaining the whole genome sequence from a P. brassicae single spore isolate. The first assembly draft showed a total length of the genome sequence of 22.8 Mb, which broadly corresponded with the previously estimation of the genome size of 18–20.3 Mb. In addition to the genome data we have also obtained transcriptome data. The exploitation of the genome will greatly advance the knowledge of this pathogen and will shed light into the evolutionary origin of P. brassicae and its biology, which will in the long run help to understand and control club root disease. Here we present some first results of the genome and transcriptome sequence analysis.
P31

ProteinConstructor - a new software for computational protein design

Grishin A, Karyagina A, Alexeevski A

Faculty of Bioengineering and Bioinformatics, Moscow State University, Russia; N.F. Gamaleya Institute for epidemiology and microbiology, Russia; A. N. Belozersky Institute for physical and chemical biology, Moscow State University, Russia

Abstract: The protein design problem can be formulated as the optimization of the protein sequence given the protein fold, function or some property, such as stability, solubility, aggregation propensity, etc. To date the scientists have used both chemical intuition and computational modeling with energy evaluation using conventional force fields and their variants to design new sequences for the known protein folds, improved and completely novel protein-protein interfaces, new enzymes and even novel protein folds. The algorithms for the computational protein design have been largely established. The most widely used conceptual scheme of the computational design is to model in 3D the various mutations and their combinations and estimate the energetic effect of these mutations. Rigid representations of protein residues called rotamers are used to sample the conformational space of the side chains, and the backbone of the protein is most often kept fixed. In spite of the apparent importance of the task and the conceptual and methodological advances, computational protein design is not yet as widely used as other computational techniques, like molecular docking or molecular dynamics simulations. The list of the software available for the molecular biologists is also limited.

Here we present the first version of the software called Protein Constructor aimed at automated design of proteins and protein complexes. The software uses the scoring function based on OPLS force field parameters for van-der-Waals energy, EEF and XlogP3 models for solvation energy evaluation, and a simple potential to account for hydrogen bonds. The scoring function was calibrated on almost 1000 point mutations and tested on a number of natural proteins and protein complexes. The software implements a number of both exact and heuristic methods to search for the best combinations of mutations, and is the first case when Simulated Bee Colony search is used in the protein design problem.
P32

Prediction of candidate cellular microRNAs targeting hepatitis B virus

Apichaya Khlaiphuengsin, Kritsada Khongnomnan, Nattanan Panjaworayan, Yong Poovorawan, Sunchai Payungporn

Department of Biochemistry, Faculty of Medicine, Chulalongkorn University, Bangkok, Thailand

Abstract: MicroRNAs (miRNAs) are a class of non-coding RNA, approximately 22 nucleotides in length with 2 nucleotides overhang on both 5′ and 3′ ends. A role of miRNAs is regulation of gene expression by triggering mRNA degradation or translational repression. Recent studies demonstrated that miRNAs also plays an important role in viral replication. In this study, we focused on bioinformatics analysis for identification of candidate human miRNAs targeting 10 genotypes (A-J) of hepatitis B virus. Candidate human miRNAs were predicted by two human miRNA prediction databases: miRBase and RNAhybrid. The miRBase was used for selection of human miRNAs with highly identical to HBV genome whereas RNAhybrid was performed for prediction of hybridization pattern. The candidate human miRNAs were analyzed based on minimum free energy (MFE) and hybridization pattern between the selected mature miRNAs and viral target genes. Each genotype of HBV was differential targeted by different miRNAs. However, there were 15 human miRNAs (mir-142-5p, mir-195, mir-214-3p, mir-384, mir-500b, mir-509-1, mir-615, mir-671-3p, mir-1273d, mir-1914, mir-3664-3p, mir-4640-5p, mir-4731-5p, mir-4999-5p and mir-5193) targeting multiple genotypes of HBV. In conclusion, this study illustrated the hybridization patterns between candidate cellular miRNAs and target viral genes among various genotypes (A-J) of hepatitis B virus by using computational analysis. The results from bioinformatics analysis provided several candidate miRNAs that needed to be confirmed by experimental analysis to ensure the effect of each predicted miRNAs. The predicted miRNAs targeting hepatitis B virus might be useful and have a potential for inhibition of hepatitis B viral replication in the future.
Pyrosequencing and T-RFLP analysis shows selective abundance among bacterial communities in metal contaminated soils

Sherlyn Tipayno, Puneet Singh Chuahan, Chang-Gi Kim, Tongmin Sa

Dept of Agricultural Chemistry, Chungbuk National University, Cheongju, Chungbuk, South Korea; 2Div. of Plant Microbe Interactions, National Botanical Research Institute, Lucknow, India; Bio-Evaluation Center, Korea Research Institute of Bioscience and Biotechnology, Cheongwon, Korea

Abstract: High-throughput sequencing analysis using Roche 454 pyrosequencing and terminal restriction length polymorphism (T-RFLP) analysis were used to compare the abundance, structure and distribution of soil bacterial communities under long-term metal and metalloid contamination. Changes in bacterial communities are essential indicators of soil quality and productivity in agricultural soils.

Soil samples were collected from three sites along the Janghang smelter in Chungman Province, representing (1) high-level contamination, (2) high-level contamination with short-term phytoremediative intervention effort, and (3) low-level contamination at an area located farther away from the source. Soil physical and chemical properties were determined and total DNA was extracted using Power Soil TM DNA isolation kit (MoBIO Lab., Inc., CA, USA). Pyrosequencing was performed by Chun’s Lab (Seoul, Korea) using the standard shotgun sequencing reagents and a 454 GS FLX Titanium sequencing System (Roche, Inc.). The 16s rRNA gene PCR amplification products were subjected to T-RFLP using the restriction enzymes Alu I, Hae III, Hha I and Msp I.

Pyrosequencing and T-RFLP revealed differences in bacterial community structure between the three sites, with distinct clustering of population in each site. T-RFLP analysis showed re-structuring was evident in highly contaminated soil, with significant reduction in abundance or loss of populations mostly belonging to Firmicutes, Bacteroides and Mollicutes. However, predominance of species belonging mostly to β-proteobacteria, including populations with key roles in the nitrogen cycle (Nitrosospira, Mesorhizobium, Azoarcus) and with known function in degradation of pollutants (Burholderia, Dechloromonas, Desulfohalobium) were seen. Pyrosequencing data showed predominance of Phylum Proteobacteria in all samples and in contrast, higher populations of γ-Proteobacteria abound in highly contaminated soil, along with populations of Acidobacteria (Groups I, 13, 18), Bacteroidoea, Spirocheates and a number of genera distributed in the different bacterial phyla.
P34

In silico approach for describing influence of different types of inhibitors on the kinetics of bisubstrate enzymatic reactions with ping-pong and sequential mechanisms

Azizyan R.A., Gevorgyan A.E., Arakelyan V.B., Gevorgyan E.S.

Yerevan State University, Department of Biophysics, Armenia

Abstract: Enzymatic reactions with simultaneous involvement of two substrates are called bisubstrate enzymatic reactions and are widely spread in different metabolic pathways. The main objective of this study is to present the results from in silico modeling of bisubstrate enzymatic reactions with ordered sequential and ping-pong mechanisms in the presence of various types of inhibitors, as well as to describe in details the inhibition effects of each type of inhibitors. Typical examples of these mechanisms are aminotransferase (transaminase) and NADH-depend dehydrogenase reactions, which play an essential role in biochemical pathway and used in clinical diagnosis. Therefore, computer-based investigation of enzyme-kinetics can highlight some useful information of specific features of these enzymes, and also data derived from in silico simulations can be helpful to find appropriate drug in appropriate quantity. In the current study we constructed several models for tree main types of inhibition: competitive, noncompetitive and uncompetitive. Models consist of set of ordinary differential equations, which makes necessary to use computational methods to understand their behavior and to be able to make predictions. To address this matter we apply numerical calculation approach for the simulation of the constructed models. We reviewed influence of each type of inhibitors in comparison with models without inhibitors. Earlier, we have shown that kinetics of these reactions without inhibitors mainly depends on ratios of local rate constants. In the case of inhibitors, we must state that the values of inhibitor concentration and inhibitor binding constant, in other words – inhibitor’s affinity, have significant contribution to the behavior of the kinetics.
Abstract: The MAMBA 7th framework EU project (http://mamba.bangor.ac.uk/) focuses on the exploration of microbes from extreme habitats, such as deep-sea hypersaline anoxic lakes (DHALs), deep-sea mud volcanoes, or hydrothermal vents. MAMBA goals are to deepen our understanding of such hitherto poorly studied or even unknown ecosystems, and to mine for enzymes with prospects in biotechnology. Here we report on two representative examples:

1) Thetis is a thalassohaline DHAL in the Eastern Mediterranean (-3,258 m), whose brine is separated from the overlying seawater by an interface that is characterized by steep salinity and oxygen gradients. A metagenomic study of the brine and the interface revealed prevalence of Epsilonproteobacteria in the interface and as yet poorly studied species from the prokaryotic KB1 and archaeal MSBL1 candidate divisions in the brine. Co-occurrence of three different carbon dioxide fixation pathways, and a ramified sulfur cycle were found, indicating that primary productions in such site thus far has been underestimated.

2) Halorhabdus tiamatea SARL4BT is a halophilic euryarchaeon that has been isolated from the Shaban Deep in the Red Sea, but Halorhabdi are also found in Mediterranean DHALs and other hypersaline habitats world-wide. H. tiamatea SARL4BT is characterized by its potential to degrade polysaccharides likely of sedimented phytoplankton origin. Complete sequencing of its genome revealed that it is among the archaea with the highest numbers of glycoside hydrolases (GHs), which makes it a particularly suited target for halotolerant GHs with industrial potential. Comparative genomics with H. utahensis AX-2T allowed for insights into the lifestyle and ecological role of this organism as well as its genus, which so far is only characterized by these two species.
Calculating binding free energies of hIFN-gamma variants and its extracellular receptor

Peicho Petkov, Elena Lilkova, Damyan Grancharov, Petko St. Petkov, Nevena Ilieva, Leandar Litov

University of Sofia "St. Kliment Ohridski", Sofia, Bulgaria

Abstract: Cytokines are crucial immunomodulating molecules which exhibit multiple biological effects by binding specific cell-surface receptors and then triggering complex intracellular signaling cascades. The interactions of cytokines and cytokine receptor have been the subject of intense research as they are involved in many important biological processes, including development, haematopoiesis, inflammation, immune responses and tissue repair and they are also associated with many diseases. In this context computational methods which allow in silico prediction of specific properties and features of cytokine-receptor interactions attract great interest.

Human interferon gamma (hIFN-g) is an important pleiotropic cytokine which is recognized by its own species-specific receptor (hIFNgR). Receptor binding sites in the hIFN-g molecule are located in three distinct areas: i) the loop between helices A and B (residues 18—26), ii) His-111, and iii) a short putative area (residues 128—131) in the flexible C-terminal domain. Apparently, the length of the C-terminal tail on hIFN-g has modulating effect on its activity, i.e. removal of up to 9 C-terminal amino acids leads to gradual increase in activity, followed by subsequent decrease in biological activity upon further deletions.

We present an implementation of a simple free energy estimation approach for calculation of binding free energies, based on Generalized Born implicit solvent (GBIS) molecular dynamics simulations. A series of GBIS MD simulations were performed to calculate binding free energies of human interferon gamma with variable chain length and its extracellular receptor. The simulation protocol was optimized and tuned to reproduce experimental data about hIFN-g – hIFNgR binding affinity.

Acknowledgements: Support from Grants DMU03/23/12.12.2011 of the Bulgarian National Science Fund is acknowledged. Calculations were performed at the Bulgarian Supercomputing Center and a local Linux cluster at the Atomic Physics Department of the Faculty of Physics of the University of Sofia “St. Kliment Ohridski”.
P37

Bioinformatic prediction of functional effects of CXC ligand 16 (CXCL16) gene polymorphisms I123T and A181V

Maja Živković1, Ljiljana Stojković1, Nevena Veljković2

1 Institute of Nuclear Sciences „Vinča”, Laboratory for Radiobiology and Molecular Genetics, University of Belgrade, Serbia, 2 Institute of Nuclear Sciences “Vinča”, Center for Multidisciplinary Research University of Belgrade, Belgrade, Serbia

Abstract: Introduction: Sequence variations of CXCL16, a chemokine which acts as a transmembrane adhesion molecule or a soluble chemoattractant for the cells expressing CXCR6, modulate its expression and/or function. We applied Informational spectrum method (ISM), Fourier transform based method for sequence analyses, to predict effects of CXCL16 polymorphisms I123T and A181V on interaction of CXCL16 with its receptor, CXCR6.

Methods: ISM transforms protein sequences into numerical series by assigning to each amino acid a value of EIIP, a parameter corresponding to long-range interaction properties. These series are further transformed into frequency domain by Fourier transform allowing for identification of dominant frequencies characteristic for protein function(s) and interaction(s). Proteins participating in mutual interaction have common dominant frequencies in their IS, which are determined by mathematical filtering in cross-spectrum (CS).

Results: CS of transmembrane CXCL16 and CXCR6 were calculated and F(0.429) was determined as characteristic frequency. We analyzed changes of amplitude (ΔA) at characteristic F (0.429) in IS of sequences corresponding to 4 CXCL16 haplotypes. Compared to the wild-type (I123A181) the other common haplotype, T123V181, had significant ΔA of −7.10%. Also, the single rare variants in haplotype with wild-type allele of another polymorphism showed the clear opposite effects (T123A181 ΔA=−13.10%; I123V181 ΔA=+5.37%).

Conclusions: Previous functional study showed that protein variants corresponding to the I123A181 and T123V181 haplotypes differed by adhesive properties. Our ISM analysis shows significantly different effects of these two haplotypes and opposite effects of the sole rare alleles, T123 and V181, on CXCL16-CXCR6 interaction. This work shows the importance of ISM analysis in prediction of haplotypes possibly responsible for changes in protein-protein interactions as guidelines for further experimental research in biomedicine.
P38

Study of effects of nanoparticles on the genes of Drosophila melanogaster through DNA microarray analysis

Arkajyoti Bhattacharya
M.Stat 2nd year student, Indian Statistical Institute, Kolkata, India

Abstract: Recently nanotechnology-based applications in many fields have increased rapidly due to the peculiar physical or chemical characteristics of engineered nanomaterials. It is extremely necessary to assess their effects upon interaction with living systems. In this context, the screening of nanomaterials to evaluate their possible effect and understand the underlying mechanisms currently represents a crucial opportunity to prevent severe harmful effects in the next future.

Result: We tested the effects of five nanoparticles on the genome of a fly. The fruit fly, popularly known as Drosophila melanogaster was fed with five different kinds of nanoparticles, namely, alumino silicate, gold, silica, silver and zinc. Also, the fly was tested in controlled situation. The nanoparticles were applied separately. After some days, the genes were observed in each case and DNA microarray experiment is done. The data in those five cases were compared with the corresponding results of those obtained in controlled situation. We did Multidimensional Scaling and hierarchical clustering and k-means clustering on the whole dataset. After that we could find out that 19 genes were statistically significantly affected by the application of each of the nano-materials. Moreover, a group of genes are also found having same molecular function.

Availability: The entire microarray was outsourced to M/S Chromous Biotech Inc, Bangalore, India from the grant money [OBT, GOI, Grant no. BT/PR9050/NNT/28/21/2007, PI: Dr. Arunava Goswami, Biological Sciences Division, Indian Statistical Institute, Kolkata]. The whole project work is done on the data obtained from the aforementioned experiment.

Keywords: Microarray-analysis, Multidimensional Scaling, Hierarchical Clustering, k-means clustering.

Usefulness: By this result, we can feel the necessity of significant efforts by the nanoscience community in designing and testing suitable nanoscale surface engineering or coating to develop biocompatible nanomaterials with no hazardous effects for human health and environment.

Biography: Arkajyoti Bhattacharya is a M.Stat 2nd year student of Indian Statistical Institute, Kolkata. He has done two summer internships under Prof. Giovanni Parmigiani (Prof. Of Harvard University). He has done 3 more projects on Nano-biotechnology.
P39

Improving SVM with Accordance Sampling in Breast Cancer Classification
A Comparison between Transductive SVM and Inductive SVM

Hala Helmi, Jonathan M. Garibaldi

School of Computer Science University of Nottingham,
Jubilee Campus, Wollaton Road, Nottingham, NG14 5AR, UK
{hqh, jmg} @cs.nott.ac.uk

Abstract: The Support Vector Machine (SVM) is a state of art classification method and it widely used in bioinformatics and other disciplines due to its high accuracy. In SVM a random sampling is used to gradually select unlabeled samples to train classifier. In this paper, we provide and explore whether using sampling method can improve SVM performance. We propose a new selective sampling method to replace the random sampling used SVM. Experimental results show that accordance-sampling method can improve SVM significantly for both types of learning for the original SVM Supervised learning and Semi Supervised learning. Supervised learning which represented in this paper as Inductive SVM using the labeled data with pre-defined classes and test data are classified into these classes too. Moreover, the other type of learning is Semi-Supervised learning which represented here as Transductive SVM, extend SVMs in that they could also treat partially labeled data. Comparing these two types of learning with accordance-sampling show that new sampling method can indeed make performance improvements over the original inductive multiclass SVM and transductive multiclass SVM.
P40

Nonlinear regression models for determination of NAD+, NADH and total NAD: application to human embryonic stem cells

A. Salykin¹, P. Kuzmic², O. Kyrylenko³, J. Musilova⁴, Z. Glatz⁴, P. Dvorak¹, S. Kyrylenko¹

1 – Department of Biology, Masaryk University, Brno, Czech Republic 2 – BioKin Ltd, Watertown, Massachusetts, USA 3 – University of Eastern Finland, Kuopio, Finland 4 – Department of Biochemistry, Masaryk University, Brno, Czech Republic

Abstract: Nicotinamide adenine dinucleotide (NAD) has been suggested as an important regulator of cellular physiology. One of the most sensitive biochemical methods for measuring the concentration of this nucleotide is enzymatic amplification. The experimental data have been conventionally analyzed by linear regression. However, the amplification process is inherently non-linear. Here we present nonlinear regression models for the precise and accurate determination of NAD+, NADH, and total NAD in eukaryotic cell lysates using a two-enzyme cycling system (alcohol dehydrogenase plus diaphorase). These models no longer require strict linearity. Instead, our method utilizes appropriate nonlinear mathematical models to analyze intrinsically nonlinear experimental data. All data analyses were performed using DynaFit software.

The nonlinear regression model for the change of resorufin fluorescence over time is a system of first order ordinary differential equations, corresponding to the Michaelis-Menten reaction mechanism. The nonlinear regression model for the calibration curve is an empirical steady-state initial rate model, involving Michaelis-Menten kinetics accompanied by substrate inhibition or activation. The accuracy of our data analysis approach was assessed by two independent methods. First, the sum of NAD+ plus NADH amounts in cell lysates was compared with the total NAD in the same lysates determined independently. Second, the cycling assay method was compared with analytical capillary electrophoresis.

We suggest that the nonlinear mathematical model presented here can bring advantage of bypassing the requirements for linearity in numerous biochemical assays.
Molecular Dynamics Investigations of Thermal Induced Conformational Changes of Bone Morphogenetic Protein-2

Iuliana Aprodu, Iuliana Banu and Adrian Istrate

Dunarea de Jos University of Galati, Romania

Abstract: The knowledge of the temperature dependent conformation of bone morphogenetic protein-2 (BMP-2) is important for predicting its bioactivity in the process of osteoblasts differentiation and proliferation while using it in bone implants and scaffolds. This work is aimed at the evaluation of thermal stability of BMP-2 at single molecule level, using Molecular Dynamics (MD) approach. The crystal structure of human BMP-2 was obtained from the Brookhaven Protein Data Bank (3BMP). MD simulations were carried out using GROMACS code. The BMP-2 model was minimized, solvated and equilibrated at temperatures resembling the room temperature (25°C) and human body conditions (37°C). The equilibrated models were further characterized in terms of conformational properties. The root mean square error obtained by least squares fitting the models equilibrated at 25°C and 37°C was 5.34Å indicating important conformational rearrangements. The temperature caused the reduction of the strands and the increase of helices content within protein structure. Regardless of temperature, the secondary structure is dominated by the strand motifs which are organized within three parallel beta sheets representing 40.6% of the protein secondary structure at 25°C and 26.4% at 37°C. Although the total number of helices formation is the same at tested temperatures, only the ~15 Å long helix (Asn59-Asn68) could be identified in both structures. At human body temperature the right-handed 3-10 helix structure disappears. The unordered secondary structure was dominated by the beta turns motifs: the total number of beta turns decreased from 13 at 25°C to 12 at 37°C, whereas the number of gamma turns increased from 1 to 7. According to our results, BMP-2 behavior is influenced by small temperature fluctuations. Short-term development will involve mechanical characterization of BMP-2.

Acknowledgements: This research has been supported by the National Research Grant PN-II-PT-PCCA-140, UEFISCDI.
Effectively computing dynamically equivalent structures of large biochemical reaction networks

János Rudan, Gábor Szederkényi, Katalin M. Hangos

Faculty of Information Technology, Pázmány Péter Catholic University; Computer and Automation Research Institute, Hungarian Academy of Sciences, Hungary

Abstract: Biochemical reaction networks (CRNs) obeying the mass action law constitute a widely accepted modeling tool to describe dynamical processes in living environments such as signalling pathways or metabolic networks. It has been known since the 1970’s that different reaction structures can produce exactly the same dynamics in the space of species concentrations. Such networks are called macro-equivalent or dynamically equivalent. This phenomenon may seriously hamper network structure inference and parameter estimation from concentration measurements [1]. It is therefore important to know whether a reaction network structure itself is sufficiently well-defined (i.e. unique) or not for the corresponding dynamics. Theoretical results and optimization tools to solve this problem set have been published recently [1]. However, most of the published computational methods were based on integer programming, where the integer variables were used to track the presence of individual reactions in the network. This caused serious computational difficulties in the case of larger networks, since integer programming is known to be NP-hard generally. This contribution presents algorithmic improvements for the circumvention of this computational problem. Using the theoretical properties of certain canonical dynamically equivalent CRN structures, fast optimization methods based on pure linear programming (without integer variables) are presented and compared for the computation of CRN structures with prescribed properties (e.g. containing the minimal or maximal number of reactions) that allow us to decide whether a structure is mathematically/practically unique or not. As the main illustrative example, the ErbB signalling pathway model [2] containing 1082 complexes and 1654 reactions is used. This work is supported by grants TÁMOP-4.2.1.B-11/2/KMR-2011-0002 and OTKA NF 104706.
P43

Fast P-value computation for characteristic motifs within sequence families

Yann Ponty, Mireille Regnier and Joanie Weaver

INRIA Amib, Ecole Polytechnique, France and MIT

Abstract: The advent of next-generation sequencing technology has renewed the need for methods addressing a de-novo discovery of recurrent patterns, or signatures, within sets of functionally-related genomic sequences. At the core of such methods, one usually finds two key elements: i) a motif exploration procedure; ii) an evaluation function, which associates a score to the motif based on its number of occurrence within the family of interest. For the latter, it has proven beneficial to consider that some motifs are intrinsically more likely to occur at random than others, depending on their complexity and their genomic context. One must therefore consider the background model to assess the significance of an observed number of occurrences. From such models, statistical estimators such as the Z-score, the E-value, or the P-value, are usually used to assess the significance of a motif.

In this work, we propose a new algorithm to compute the significance of observing $X$ occurrences of a motif $H$ within a sequence family of interest. It supports Markov models of any order as background models, which can be automatically built from a set of sequences. Building on the recently introduced overlap graph for $H$, it computes the probability of observing $H$ within each of the sequence in the observed family. These probabilities are then combined into an exact computation of the P-value, i.e. the probability of observing more than $X$ occurrences by chance, using a dynamic-programming algorithm proposed by Bourdon, et al. Its complete complexity grows like $O(M.X + N.(|H|m + A^{k+1}))$, where $M$ is the number of sequence in the observed family, $m$ is the length of characteristic motifs, $X$ is the observed number of occurrences within the family, $N$ is the maximum length of a sequence, $A$ is the alphabet length (e.g. $A = 4$ for RNA/DNA), and $k$ is the order of the background Markov model.

The algorithm was implemented within the RogueProb software. Written using the Java language, the algorithm can be run on any platform, or through a web browser. Additional features include the computation of the Markov model from a, possibly large, collection of files. The software can be tested, and downloaded freely on a dedicated website at: http://www.lix.polytechnique.fr/bioinfo/RogueProb
Combined T-cell repertoire analysis by NGS with mathematical modeling of viral dynamics to optimize treatment of BK virus post kidney transplantation

Ulrik Stervbo ¹, Leon Kuchenbecker ¹, Mikalai Dzubianau ¹, Thomas Schachtner ¹, Nina Babel ¹,², Petra Reinke ¹,², and Avidan U. Neumann ¹,³

¹) Berlin-Brandenburg Center for Regenerative Therapies (BCRT) and ²) Department of Nephrology, Charite University Medicine Berlin, Berlin, Germany ³) Goodman Faculty of Life Sciences, Bar-Ilan University, Ramat-Gan, Israel.

Abstract: Reactivation of polyoma BK virus (BKV) post kidney transplantation poses a serious threat to patient and graft. While in some patients BKV reactivation is transient, in others viremia is persistent. The correlates of immunity and therapy protocol necessary to successfully control the virus are not yet clear. Deep sequencing by NGS technology allows to obtain extensive information on the T-cell repertoire repertoire. Mathematical modeling of viral dynamics had achieved important clinical results in HIV, HCV and HBV (Neumann et al, Science 1998). Here, we used deep sequencing of the CDR3 receptor by NGS technology to analyze the T-cell repertoire in correlation with IFN-gamma specific immune response (ELISPOT) measurements and mathematical modeling of the BKV dynamics in 12 patients with frequent sampling during BKV reactivation post kidney transplantation. Fitting of the combined immunological and virological data indicated 2 waves of immune response against different viral antigens with different functional properties. The first response against the structural BKV-VP antigens started only after a switch in immune suppression, and gave rise to limited viral decline by blocking viral replication. The second response against the non-structural BKV-T antigens gave rise to viral clearance by killing of infected cells. Deep sequencing of the T-cell repertoire allowed the identification and quantification of individual T-cell clones kinetics. Furthermore, analysis of T-cell repertoire in different tissues allowed to explain why in some patients the immune response was not effective.

In conclusion, modeling the combined viral and immune kinetics was successful in elucidating the dynamical properties of different T-cell specificities. NGS technology shows promising results to better characterize the immune repertoire in correlation with viral dynamics. Our results indicate possible protocols to allow optimizations of the immune suppression regimen post transplantation.
P45

A phylogenetics system recommendation framework for accurate tree reconstruction and analyses

Ahmed Halioui, Petko Valtchev, Abdoulaye Baniré Diallo

Computer Science dept, Université du Québec à Montréal, Canada

Abstract: Molecular phylogenetic and evolution describe a specific vocabulary, a combination of expressions, proper to biology and genetics. The amount of genomic information increases dramatically as the crowdsourcing diverse phylogenetic methods and approaches. This makes phylogenetic analysis a more complex and a complicated task. Such a mosaic could confuse an expert while modeling his phylogenetic inferences. Here, we propose a phylogenetic recommendation system providing “standard” analysis protocols accepted by the community of phylogeny. Our system is based on text mining and presenting an ontology of phylogenetic practices. A phylogenetic ontology represents concepts, as approaches and methods, and properties, as constraints on the way an expert should distinguish standard techniques for a given problematic. Hidden practices (patterns) are rediscovered from the phylogenetic literature and model as phylogenetic workflows using our developed pipeline (Armadillo 1.1). It is important to notice that our approach to ontology learning provides a broad coverage of the phylogenetic domain. Learning new patterns from the ontology provides maintenance and reliability to the phylogenetic knowledge representation, This is the core function of our recommender system. The lack of full knowledge about phylogenetic reality manifests itself an ambiguity and nondeterminism. Currently, a small set of the ontology is created. Its maintenance and validation are provided in a semi-automatic way. Recommendations are formally expressed through these patterns.
P46

Using an automated image-analysis platform to interpret the results from the screening of an Arabidopsis mutant collection on reactive oxygen species-eliciting chemicals

Veselin Petrov¹, Jos Schippers²,³, Ivan Minkov¹, Bernd Mueller-Roeber²,³, Tsanko Gechev¹

¹ Department of Plant Physiology and Molecular Biology, University of Plovdiv, 24 Tsar Assen str., Plovdiv 4000, Bulgaria ²Max Planck Institute of Molecular Plant Physiology, Am Muehlenberg 1, 14476 Potsdam, Germany ³Institute of Biochemistry and Biology, University of Potsdam, Karl-Liebknecht-Str. 24–25, 14476 Potsdam, Germany

Abstract: In order to unveil potential novel players of the ROS-regulatory network, a population of ~2,000 randomly selected Arabidopsis thaliana T-DNA insertion mutants was screened for ROS sensitivity/resistance by growing seedlings on medium supplemented with stress-inducing concentrations of the superoxide-producing herbicide methyl viologen or the catalase inhibitor 3-amino-triazole. Mutant phenotypes were assessed using a Lemnatec automated image-processing platform and subsequent analysis of the photographs with a user-made algorithm created in the LemnaGrid software environment. Seedlings were characterized by two parameters, the total object area (in pixel²) and the ratio between healthy(green)/chlorotic(yellow) tissues, which provided quantitative information to interpret the results and verify the data. After rescreening the first batch of 33 identified candidates, the ROS-related phenotype was confirmed for three of them: SALK_099042, SALK_020158 and SALK_113517, which are affected by T-DNA insertions in genes encoding a Ring/Ubox superfamily protein, ABI5 binding protein 1 (AFP1), and a protein of unknown function, respectively.
Identification of miRNAs and their isomiR sequences in human uterine leiomyoma.

I. Milev, B. Georgieva, G. Minkov, V. Baev

University of Plovdiv, Dept. Plant Physiology and Molecular Biology, Bulgaria

High-throughput sequencing or deep sequencing has emerged as a direct small RNA examination method. Profiling of microRNAs by deep sequencing allows the discovery of novel miRNAs, measuring absolute abundance as well as detection of variants from “reference” mature sequences that are called “isomiRs”. Deep sequencing generates millions of small RNA sequence reads from a given sample profiling and provides a framework for exploring the transcriptional complexity. However, high-throughput sequencing data sets present new problems related to data processing and extraction of all information from the huge amount of obtained data.

miRNAs play important roles in cellular control and in various disease states such as cancers. Identifying the whole repertoire of miRNAs and understanding their expression patterns is therefore an important goal.

Uterine leiomyomas are benign neoplasms of the smooth muscle cells of the uterus and are the most common tumors in women. Although they are not malignant, these tumors may cause significant reproductive and gynecological problems, such as menorrhagia, dysmenorrhoea, chronic pelvic pain, as well as infertility and pre-term delivery. However, there is little understanding of the processes of leiomyomas growth and regression at the cellular and molecular levels.

Here, we present a new web-based tool for identification of miRNA, their cognate isomiRs and investigation of differentially expressed miRNAs using NGS datasets. The software also detects isomiRs of a miRNA with higher copy number relative to their mature reference sequence indexed in mirBase.
From starPRO database to transposon indel polymorphism in promoters in natural variants of *Arabidopsis*

Naydenov Mladen, Anachkov Nikolay, Milev Ivan, Apostolova Elena, Vachev Tihomir, Baev Vesselin and Yahubyan Galina

**Dept. Plant Physiology and Molecular Biology, University of Plovdiv, Bulgaria**

**Abstract:** RNA-dependent DNA methylation (RdDM) is an important regulatory event involved in repressive epigenetic modifications that can trigger transcriptional gene silencing. The criteria we used to pick out promoter sequences targeted by RdDM in *Arabidopsis thaliana* were the main RdDM hallmark properties: 24nt siRNAs as inducers of DNA methylation and transposable elements (TE) as one of the major targets of RdDM. Those genes whose promoters comprised overlapping sites for 24nt siRNA hits, TE and DNA methylation, were defined as candidates that might be silenced by RdDM. On this basis, a publicly available 24nt siRNA-centered database called starPRO was developed that allows users easily to discover whether a particular promoter sequence is related to the RdDM-associated features.

Since the mechanisms of epigenetic regulation are responsible for the formation of heritable epigenetic gene variants (epialleles) and also regulate TE mobility, we selected several natural variants (ecotypes) of *A. thaliana* in order to study the primary structure of promoters of stress-responsive genes that were identified with starPRO as putative targets for RdDM silencing. Using end-point PCR and sequencing for the analysis, we discovered insertion-deletion (indel) polymorphism for several promoters - in some accessions the TE was found deleted, in others it was found inserted, and there was a part of promoters with no changes in the sequence context. Our findings reveal that different environmental conditions may contribute to a particular TE profile in *cis-acting* regulatory sequences such as promoters.
Sequence analysis of sRNA population recovered from PSTVd infected O.ramosa

Desislava Ivanova, Vesselin Baev, Ivan Minkov, Mariyana Gozmanova

Dept. Plant Physiology and Molecular Biology, University of Plovdiv, Bulgaria

Abstract: Viroids are the smallest (246-401nt) circular single stranded RNAs among the known plant pathogens, not coding for any protein, able to self-replicate and induce disease in higher plants.

PSTVd is the first identified viroid species, naturally infecting Solanaceae plants. Recently the experimental host range of PSTVd was extended with the parasitic plant O.ramosa. Similarly to higher plants non-photosynthetic parasitic plants encode, process and accumulate short 20-24nt RNAs (sRNAs) referred to as micro RNAs (miRNAs) and small interfering RNAs (siRNAs) that combat with biotic stress.

In order to reveal the presence of sRNAs specific to PSTVd infection and to define conserved plant miRNA families in the genome of O.ramosa we performed high-throughput sequence analysis of sRNA pull recovered from nonsymptomatic O.ramosa stem. We report PSTVd specific siRNAs in infected O.ramosa and lack of such in non-infected O.ramosa. Further, we mapped siRNAs derived from PSTVd upon viroid genome sequence and thus we were able to locate their origin and point the hot spots. Additionally we found several differentially expressed miRNA classes in infected O.ramosa, suggesting that these small RNAs may play a defensive role in response to PSTVd.
P50

Prediction of secondary structure of chimaeric viral proteins

Milena Mazalovska and Gergana Zahmanova

Department Plant Physiology and Molecular Biology, University of Plovdiv, Bulgaria

Abstract: Virus-like particles (VLPs) consist of viral capsid proteins that can initiate an immune response but do not include the genetic material require for viral replication. A variety of VLPs have been developed for use as a vaccines and epitope platforms. Chimaeric viral constructs have been developed by insertion of M2e peptide from Avian influenza into major immunodominant loop of Hepatitis B core protein and Woodchuck hepatitis core protein. By linking M2e peptide (the extracellular domain of influenza M2 protein) to viral core protein the sequences can be presented in a much more immunogenic form to the immune system. The models of chimaeric viral proteins have been predicted using the program SWISS-MODEL Workspace. SWISS-MODEL is a fully automated protein structure homology-modeling server (Arnold K., et all 2006). To validate the results from protein structure prediction the sequences encoding the chimaeric molecules have been inserted into a CPMV-HT vector, pEAQ-HT, which has been created at JIC (Lomonossoff G., et al. 2009). The constructs have been transformed into Agrobacterium tumefaciens and bacterial suspensions infiltrated into Nicotiana benthamiana. At various days post-infiltration (DPI), samples have been collected and the proteins extracted using standard procedures. Expression of the viral chimaeras have been examined by SDS PAGE of the samples, using Coomassie staining, Western blot and TEM.
P51

Oxidative stress provokes distinct transcriptional responses in the stress-tolerant atr7 and stress-sensitive loh2 Arabidopsis thaliana mutants as revealed by multi-parallel quantitative real-time PCR analysis of ROS marker and antioxidant genes.

Mehterov N1, Balazadeh S2,3, Hille J4, Toneva V1, Mueller-Roeber B2,3, Gechev T1

1 Department of Plant Physiology and Molecular Biology, University of Plovdiv, 24 Tsar Assen str., Plovdiv 4000, Bulgaria 2Max Planck Institute of Molecular Plant Physiology, Am Muehlenberg 1, 14476 Potsdam, Germany 3Institute of Biochemistry and Biology, University of Potsdam, Karl-Liebknecht-Str. 24–25, 14476 Potsdam, Germany, 4 Department Molecular Biology of Plants, Groningen Biomolecular Sciences and Biotechnology Institute (GBB), University of Groningen, Nijenborgh 7, 9747 AG, Groningen, The Netherlands

Abstract: The Arabidopsis thaliana atr7 mutant is tolerant to oxidative stress induced by paraquat (PQ) or the catalase inhibitor aminotriazole (AT), while its original background loh2 and wild-type plants are sensitive. Both, AT and PQ, which stimulate the intracellular formation of H2O2 or superoxide anions, respectively, trigger cell death in loh2 but do not lead to visible damage in atr7. To study gene expression during oxidative stress and ROS-induced programmed cell death, two platforms for multi-parallel quantitative real-time PCR (qRT-PCR) analysis of 217 antioxidant and 180 ROS marker genes were employed. Primers for the individual genes were designed using QuantPrime design tool and the efficiencies of the polymerase chain reactions were estimated using the LinRegPCR software. Subsequent data analysis was performed using SDS 2.2.1 software (Applied Biosystems) and relative mRNA abundance was calculated using the comparative 2-ΔΔCt method and normalized to the corresponding reference gene levels. Finally, resulting data sets were log transformed and visualized by MEV (MultiExperiment Viewer)- created heat maps. The qRT-PCR analyses revealed AT- and PQ-induced expression of many ROS-responsive genes mainly in loh2, confirming that an oxidative burst plays a role in the activation of the cell death in this mutant. Some of the genes were specifically regulated by either AT or PQ, serving as markers for particular types of ROS. Genes significantly induced by both AT and PQ in loh2 included transcription factors (ANAC042/JUB1, ANAC102, DREB19, HSFA2, RRTF1, ZAT10, ZAT12, ethylene-responsive factors), signaling compounds, ferritins, alternative oxidases, and antioxidant enzymes. Many of these genes were upregulated in atr7 compared to loh2 under non-stress conditions at the first time point, indicating that higher basal levels of ROS and higher antioxidant capacity in atr7 are responsible for the enhanced tolerance to oxidative stress and suggesting a possible tolerance against multiple stresses of this mutant.
PS2

FCA as a means for consolidation of clustering results derived from multiple experiments

Veselka Boeva\(^1\), Anna Hristoskova\(^2\), Elena Tsiporkova\(^3\), Elena Kostadinova\(^1\)

\(^1\)Computer Systems and Technologies Department, Technical University of Sofia-branch Plovdiv Tsanko Dyustabanov 25, 4000 Plovdiv, Bulgaria vboeva@tu-plovdiv.bg, \(^2\)Department of Information Technology Ghent University - IBBT Gaston Crommenlaan 8 (201), 9050 Ghent, Belgium anna.hristoskova@intec.UGent.be, \(^3\)Sirris, ICT and Software Engineering group The Collective Center for the Belgian technological industry Brussels, Belgium elena.tsiporkova@sirris.be

Abstract: Gene clustering is one of the most important top-down microarray analysis techniques when it comes to extracting meaningful information from gene expression profiles. Clustering algorithms are used to divide genes into groups according to the degree of their expression similarity. Such a grouping indicates that the respective genes are correlated and/or co-regulated, and subsequently indicates that the genes could possibly share a common biological role.

Presently, with the increasing number and complexity of the available gene expression data sets the combination of data from multiple microarray studies addressing a similar biological question is gaining high importance. The integration and evaluation of multiple datasets yields more reliable and robust results since they are based on a larger number of samples and the effects of the individual study-specific biases are diminished. One useful way for integration analysis of the data from different experiments is to aggregate their clustering results into a consensus clustering which both emphasizes the common organization in all the datasets and at the same time reveals the significant differences among them.

In this work, we examine and demonstrate the potential of Formal Concept Analysis (FCA) for consolidation and analysis of clustering results derived separately from a set of microarray experiments studying the same biological phenomenon. We consider two approaches to consensus clustering of gene expression data across multiple experiments. The first algorithm consists of two distinctive steps: 1) a preliminary selected clustering algorithm (e.g. k-means) is initially applied to each experiment separately, which produces a list of different clustering solutions, one per experiment; 2) these clustering solutions are further transformed into a single clustering result by employing FCA, which allows to analyze and extract valuable insights from the data. In the second algorithm, the available microarray experiments are initially divided into groups of related datasets with respect to a predefined criterion (e.g. experimental settings). The rationale behind this is that if experiments are closely related to one another, then these experiments may produce more accurate and robust clustering solution. Subsequently, the Particle
Swarm Optimization (PSO)-based clustering algorithm is applied to each group of experiments separately. The result is a list of different clustering solutions, one for each group. These clustering solutions are pooled together and again analyzed by employing FCA. Notice that FCA produces a concept lattice where each concept represents a subset of genes that belong to a number of clusters. In both algorithms the generated concepts compose the final clustering partition.

The foregoing clustering approaches have been demonstrated to have certain advantages with respect to the traditional consensus clustering techniques, namely both methods: 1) use all data by allowing potentially each experiment (or group of related experiments) to have a different set of genes, i.e. the total set of studied genes is not restricted to those contained into all datasets; 2) are better tuned to each experimental condition by identifying the initial number of clusters for each experiment (or group of related experiments) separately depending on the number, composition and quality of the gene profiles; 3) avoid the problem with ties (i.e. a case when a gene is randomly assigned to a cluster because it belongs to more than one cluster) by employing FCA in order to analyze together all the partitioning results and find the final clustering solution representative of the whole experimental compendium.
P53

Detecting microbial pathogens in next generation sequencing data

Nikolay Nikolov, Nikolay Borisov, Tomislav Ilicic, Simon Watson, Matt Cotton, Astrid Gall, Anne Palser, Paul Kellam

Wellcome Trust Sanger Institute, UK, Aberystwyth University, University of Applied Sciences Upper Austria, Austria

Abstract: The genomic material of microbial pathogens is present in very small quantities in samples taken from their hosts. We present a bioinformatics pipeline that performs consecutive rounds of QC, host genome elimination, de novo assembly and taxonomic classification. Using this pipeline we have been able to detect microbial presence of 0.2pg in a 2mcg host sample (concentration of 1x10^-7) from next generation sequencing data.

P54

Clustering of Microarray Data Analysis

Cigdem Erol, Serra Celik, Murat Gezer, Sevinc Gulsecen

Istanbul University Informatics Department, Turkey

Abstract: Nowadays microarray methods are often used to investigate of genome of an organism. For this technology, that allows the measurement of expression levels of thousands of genes, interdisciplinary collaboration is required to reveal biologically meaningful information from huge amount of expression data. Data mining is an essential technology and clustering algorithms are especially used in analysis of gene expression levels. Instead of choosing an algorithm randomly, different clustering algorithms should be applied to data. After deciding on algorithm other stages of data mining can be applied for achieving accurate and reliable information. In this study, by applying different algorithm to a data set from published in NCBI GEO database. We wil try to find the proper algorithm and than proceed on datamining.
An approach to a metagenomic data processing workflow

Milko Krachunov¹, Dimitar Vassilev²

¹- Faculty of Mathematics and Informatics, Sofia University, "St. Kliment Ohridski", Sofia, Bulgaria, ²- Bioinformatics group, AgroBioInstitute, Sofia, Bulgaria

Abstract: Metagenomics is a rapidly growing field that focuses on the study of genetic material collected from largely unexplored heterogeneous biological environments, which has been greatly driven by the ongoing advancements in high-throughput sequencing technologies. As a result, both the data preparation and the subsequent in silico experiments pose unsolved technical and theoretical challenges, as there aren't any well-established approaches, and new expertise and software are constantly emerging.

Due to the nature of the data obtained – which presently lacks any inherent reference points or a standard for validation – every study involves the computational challenges associated with de novo sequencing, which are further exacerbated by the need to deal with a larger degree of uncertainty, significantly larger data and the need to adapt the data processing to your particular experiment, often multiple times.

Our project initial goal is the development and the examination of an approach for error detection and correction aiming to reduce some of the difficulty caused by the lack of references. It relies on the calculation of weighted frequencies based on the local similarity between sequence pairs. In a metagenomic sample, you can't avoid the appearance of related organisms that inevitably lead to the introduction of noise in the data. A carefully defined noise filter can reduce their influence and or even take advantage of their presence as a confirming factor for very well conserved orthologous regions.

During the course of the work it became clear that a tool or a library for managing, running and distributing the processing components would greatly reduce the amount of manual work required to run any metagenomic experiment, and the error detection became a part of a larger project to create a fully-asynchronous library for performing genomic processing, capable of running arbitrary external tools, and providing a general command-line tool for launching preconfigured workflows.

By using an asynchronous network framework and creating interfaces fully adhering to its idioms, the tasks of scheduling, parallelising and distributing the jobs across networked computers becomes easy. It also becomes easier to run a variety of components in variety of combinations while doing basic resource management. Such approach would also facilitate the running and comparison of different error detection approaches with different data processing layers.
List of Participants and Poster number

P4
Abdoulaye Banire Diallo
Université du Québec à Montréal, Canada, Computer Science
PO BOX 888 Downtown station, Montreal, Canada
e-mail: diallo.abdoulaye@uqam.ca

P31
Alexander Grishin
Moscow State University, Faculty of Bioengineering and Bioinformatics
Faculty of Bioengineering and Bioinformatics, office 433, Moscow, Russian Federation
e-mail: grishin@kodomo.fbb.msu.ru

P29
Anastasia Slustikova Lebedik
Tomas Bata University in Zlin, Department of Informatics and Artificial Intelligence
Czech Republic
e-mail: lebedik@fai.utb.cz

Andrey Kajava
Centre de Recherches de Biochimie Macromoléculaire, CNRS Université Montpellier, France
e-mail: andrey.kajava@crbm.cnrs.fr

P13
Andrzej Lyskowski
ACIB - Austrian Centre of Industrial Biotechnology
IMB - Institute of Molecular Biosciences, University of Graz
Humboldtstrasse50,III/194, Graz, Austria
e-mail: andrzej.lyskowski@uni-graz.at

P40
Anton Salykin
Masaryk University, Department of Biology
Kamenice 5, b. A3, Brno, Czech Republic
e-mail: a.salykin@gmail.com

P38
Arkajyoti Bhattacharya
Indian Statistical Institute, Kolkata, e-mail: arkajyoti.bhattacharya@gmail.com

P28
Assyl Bari
Al-Farabi Kazakh National University, Department of Biology and Biotechnology
71 Al-Farabi av., Almaty, Kazakhstan
e-mail: assyl.bari@gmail.com

P44
Avidan Neumann
Berlin-Brandenburg Center for Regenerative Therapies (BCRT) and Institute for Theoretical Biology (ITB), Humboldt University
e-mail: auneumann@gmail.com

P20
Benjamin Kumwenda
University of Pretoria, Biochemistry, Bioinformatics and Computational Biology Unit
3-3 FABI Square, Lynwood Road, Pretoria, South Africa
e-mail: benjkum@gmail.com

Boryana Georgieva
University of Plovdiv, Plant physiology
and Molecular biology  
Tsar Assen Str. 24, Plovdiv, Bulgaria  
e-mail: boriana_ge@abv.bg

P24  
Chiara Broccanello  
DAFNAE, University of Padova, Italy  
Viale UniversitA 16, Legnaro (PD), Italy  
e-mail: stevanato@unipd.it

P27  
Chun-Wei Tung  
Kaohsiung Medical University, School of Pharmacy  
100, Shih-Chuan 1st Road, Kaohsiung, 80708, Taiwan, Kaohsiung, Taiwan  
e-mail: cwtung@kmu.edu.tw

P54  
Cigdem Erol  
Istanbul University, Informatics, Istanbul Universitesi Enformatik  
Bolumu, Vezneciler  
Istanbul, Turkey  
e-mail: sevincg@yahoo.com

P49  
Desislava Ivanova  
University of Plovdiv, Plant physiology and Molecular biology  
Tsar Assen Str. 24, Plovdiv, Bulgaria  
e-mail: desi_plamenova@abv.bg

P55  
Dimitar Vassilev  
Bioinformatics group, AgroBioInstitute, Sofia, Bulgaria  
e-mail: jim6329@gmail.com

Dimitrina Konstantinova  
Medical University Varna, Medical Genetics  
1 Hristo Smirnensky Str, Varna, Bulgaria  
e-mail: dimitrina1956@yahoo.com

P7  
Dominique Chu  
University of Kent, School of Computing  
Cornwallis, Canterbury, United Kingdom  
e-mail: d.f.chu@kent.ac.uk

Dorothea Kazazi  
Hellenic Pasteur Institute, Department of Microbiology  
127 Vasilissis Sofias avenue, TK 11521, Athens, Greece  
e-mail: dorothea.kazazi@gmail.com

P8, P36  
Elena Lilkova  
University of Sofia St. Kliment Ohridski, Faculty of Physics  
5 J. Bouchier Blvd., Sofia, Bulgaria  
e-mail: elilkova@phys.uni-sofia.bg

P14  
Elisa Cilia  
Université Libre de Bruxelles, Département d’Informatique  
Boulevard du Triomphe - CP 212  
Brussels, Belgium  
e-mail: ecilia@ulb.ac.be

Enrique Morett  
Universidad Nacional Autonoma de Mexico, Instituto de Biotecnologia  
Av. Universidad 2001 Chamilpa  
Cuernavaca, Mor., Mexico  
e-mail: emorett@ibt.unam.mx

P3  
Anna Faddeeva-Vakhrusheva  
VU Amsterdam, Animal Ecology  
De Boelelaan 1085, Amsterdam, The Netherlands  
e-mail: avakhrusheva@gmail.com
P1, P24
Gaurav Sablock
Research and Innovation Center
Environmental Biotechnology Platform
Fondazione Edmund Mach (FEM)
Istituto Agrario San Michele (IASMA)
e-mail: sablokg@gmail.com

P12, P13
Georg Steinkellner
ACIB - Austrian Centre of Industrial Biotechnology, IMB - Institute of Molecular Biosciences, University of Graz, Humboldtstraße 50, III/194, Graz, Austria
e-mail: georg.steinkellner@uni-graz.at

P9, P10
Gianfranco Politano
Politecnico di Torino, Control and Computer Engineering
Corso Duca degli Abruzzi 24, Torino, Italy
e-mail: gianfranco.politano@polito.it

Giovanna Capovilla
DAFNAE, University of Padova, viale dell’Università 16, 35020 Legnaro (Padova), Italy
e-mail: stevanato@unipd.it

P39
Hala Helmi
Nottingham University, Computer Science
703 Marco Island Huntingdon st, Nottingham, UK
e-mail: hala.helmi@hotmail.com

P26
Helena Petrosova
Masaryk University, Department of Biology
Kamenice 5, Building A6, Brno, Czech Republic
e-mail: hpetros@med.muni.cz

Hristo Yordanov
ACM2
e-mail: cjordanov@acm2.com

P41
Iuliana Aprodu
Dunarea de Jos University of Galati, Food Science and Engineering
Domneasca 111, Cod 800201, Galati, Romania
e-mail: iuliana.aprodu@ugal.ro

P48, P47
Ivan Milev
University of Plovdiv, Plant physiology and Molecular biology
Tsar Assen Str. 24, Plovdiv, Bulgaria
e-mail: milev@plantgene.eu

Ivona Uzunova
Chrono
e-mail: ivona.uzunova@chrono-bg.com

P42
János Rudan
Pázmány Péter Catholic University, Faculty of Information Technology
Práter utca 50/a., Budapest, Hungary
e-mail: rudan.janos@itk.ppke.hu

P30
Johan Fogelqvist
Swedish University of Agricultural Sciences, Department of Plant Biology and Forest Genetics, PO-Box 7080, SE-75007 Uppsala, Sweden
e-mail: johan.fogelqvist@slu.se

P35
Johannes Werner
Max Planck Institute for Marine Microbiology, Microbial Genomics and Bioinformatics Group
Celsiusstrasse 1, D-28359 Bremen, Germany
e-mail: jwerner@mpi-bremen.de
José R. Valverde  
Centro Nacional de Biotecnología,  
CSIC, Spain  
e-mail:

P17  
Katarzyna Walczewska-Szewc,  
University of Gdańsk, Faculty of  
Mathematics, Physics and Informatics  
Wita Stwosza 57, Gdańsk, Poland  
e-mail: fizkws@ug.edu.pl

Klaas Vandepoele  
VIB / Ghent University, Belgium  
e-mail: Klaas.Vandepoele@psb.vib-  
ugent.be

P15, P16  
Lorentz JÂNTSCHI  
University of Agricultural Science and  
Veterinary Medicine Cluj-Napoca  
3-5 Calea Mănăștur, 400372 Cluj-  
Napoca, Romania.  
e-mail: lorentz.jantschi@gmail.com

P37  
Maja Zivkovic  
Institute of Nuclear Sciences Vinca,  
Laboratory for Radiobiology and  
Molecular Genetics  
Mike Petrovica Alasa 12-14, P.O.Box  
522, Belgrade, Serbia  
e-mail: majaz@vinca.rs

Mario Fares  
University of Dublin, Ireland  
e-mail: FARESM@tcd.ie

P4  
Marisa Fabiana Nicolás  
LNCC, Labinfo  
Av Getulio Vargas 333 Zip Code  
25651-075, Petrópolis RJ, Brazil  
e-mail: marunicolas@gmail.com

Mary Hachmerian  
Medical Genetics, Medical University  
Varna  
1 Hristo Smirnensky Str, Varna,  
Bulgaria  
e-mail: mary.hachmeriqn@abv.bg

P1  
Martina Popova  
University of Plovdiv, Plant physiology  
and Molecular biology  
Tsar Assen Str. 24, Plovdiv, Bulgaria  
e-mail: marti8811@gmail.com

P50  
Milena Mazalovska  
University of Plovdiv, Plant Physiology  
and Molecular Biology  
Tsar Assen str. 24, Plovdiv, Bulgaria  
e-mail: mazalovska@yahoo.com

Milka Georgieva  
Dept of Pathology, Laboratory of  
molecular pathology, University  
Hospital St. Marina  
Hristo Smirnenski str 1, Varna 9010,  
Bulgaria  
e-mail: milkageorgieva@yahoo.com

P48  
Mladen Naidenov  
University of Plovdiv, Plant physiology  
and Molecular biology  
Tsar Assen Str. 24, Plovdiv, Bulgaria  
e-mail: mldlefrancais@abv.bg

Nadia Dimitrova  
Diamed Ltd, Agilent Technologies  
e-mail: life_sciences@diamed.bg

P21  
Natalia Mitiushkina  
N.N. Petrov Institute of Oncology,  
Laboratory of Molecular Oncology  
Pesochny-2, 197758, Leningradskaya,  
68, St.-Peterburg, Russia  
e-mail: nmmail@inbox.ru
20-22 September, Varna, Bulgaria

P4
Nicholas Costa Barroso Lima
LNCC, Labinfo
Av Getulio Vargas 333 Zip Code
25651-075, Petrópolis RJ, Brazil
e-mail: nicholas@lncc.br

P51
Nikolay Mehterov
University of Plovdiv, Plant physiology and Molecular biology
Tsar Assen Str. 24, Plovdiv, Bulgaria
e-mail: nikolaymehterov@uni-plovdiv.bg

P53
Nikolay Nikolov
Wellcome Trust Sanger Institute, Virus Genomics Group,
Wellcome Trust Genome Campus,
Hinxton, UK
e-mail: nn1@sanger.ac.uk

P18, P19, P20
Oleg reva
University of Pretoria, Biochemistry, Bioinformatics and Computational Biology Unit
Hillcrest, Lynnwood Rd, Pretoria, South Africa
e-mail: oleg.reva@up.ac.za

P11
Pedro Ballester
EMBL-EBI, Structural Bioinformatics
Wellcome Trust Genome Campus,
Cambridge, UK
e-mail: pedro.ballester@ebi.ac.uk

P8, P36
Peicho Petkov
University of Sofia St. Kliment Ohridski, Faculty of Physics, Atomic Physics Department

5, James Bouchier, Blvd., Sofia, Bulgaria
e-mail: peicho@phys.uni-sofia.bg

Petio Petkov
LKB
e-mail: p.petkov@lkb.eu

P34
Rafayel Azizyan
Yerevan State University, Department of Biophysics
29 apt., 9 Nalbandyan Str., 0010, Yerevan, Yerevan, Armenia
e-mail: rafayel.azizyan@gmail.com;

P19
Rian Ewald Pierneef
University of Pretoria, Biochemistry, Bioinformatics unit
Lynnwood Road, Pretoria, South Africa
e-mail: repierneef@live.com

P22
Rumyana Dodova
Molecular Medicine Center,
Department of Medical Chemistry and Biochemistry
Medical University Sofia, Bulgaria
e-mail: r.dodova@mmcbg.org

Ruth Wagner
Monsanto
Leeuwenhoekweg 52, Bergschenhoek, Netherlands
e-mail: ruth.wagner@monsanto.com

Ruud de Maagd
Business Unit Bioscience, Plant Research International, 6700 AP Wageningen, The Netherlands, and Centre for BioSystems Genomics (CBSG), 6700 AB Wageningen, the Netherlands
e-mail: ruud.demaagd@wur.nl
Samir Naimov  
University of Plovdiv, Plant physiology and Molecular biology  
Tsar Assen Str. 24, Plovdiv, Bulgaria  
e-mail: naimov0.uni-plovdiv.bg

Sebastien Briois  
Syngenta Seeds  
12 chemin de l'Hobit, F-31790 Saint-Sauveur, France  
e-mail: sebastien.briois@syngenta.com

Sevinc Gulsecen  
Istanbul University, Informatics  
Istanbul Universitesi Enformatik Bolumu, Vezneciler, Istanbul, Turkey  
e-mail: sevincg@yahoo.com

Sherlyn Tipayno  
Chungbuk National University, Department of Environmental and Biological Chemistry  
Gaesindong, Heungdok-gu, Cheongju City, Chungbuk, South Korea  
e-mail: amygdala_ph@yahoo.com

Simon Schliesky  
Heinrich Heine University, Institute of Plant Biochemistry  
Universitätsstr. 1, Düsseldorf, Germany  
e-mail: simon.schliesky@uni-duesseldorf.de

Sorana D. BOLBOACA  
Iuliu Hatieganu University of Medicine and Pharmacy Cluj-Napoca, Medical Informatics and Biostatistics  
6 Louis Pasteur, Cluj-Napoca, Romania  
e-mail: sbolboaca@umfcluj.ro

Veselin Petrov  
University of Plovdiv, Plant Physiology and Molecular Biology, Bulgaria  
e-mail: vpetrov@plantgene.eu, vepetrov@uni-plovdiv.bg

Veselka Boeva  
Technical University of Sofia - branch Plovdiv, Computer Systems and Technologies Department  
Tsanko Dyustabanov 25, Plovdiv, Bulgaria  
e-mail: vboeva@tu-plovdiv.bg

Vesselin Baev  
University of Plovdiv, Plant physiology and Molecular biology  
Tsar Assen Str. 24, Plovdiv, Bulgaria  
e-mail: vebaev@gmail.com

Vitalii Akimenko  
Taras Shevchenko National University of Kyiv, Cybernetics  
Volodymyrska 64, 01030 Kyiv, Ukraine  
e-mail: akm15@meta.ua

Wojciech Karlowski  
Institute for Molecular Biology and Biotechnology, Poznan, Poland  
e-mail: Wojciech.Karlowski@amu.edu.pl
P5
Yuval Itan
The Rockefeller University,
Human Genetics of Infectious Diseases
1230 York Avenue, New York, USA
e-mail: yitan@rockefeller.edu
Diamed LTD – the official representative of Agilent Technologies for Bulgaria

Agilent understands the importance of being your partner in genomic research. By providing a comprehensive suite of high quality genomics arrays, reagents, instruments and software, your work is advanced with flexibility, sensitivity and specificity vital to answering biological questions. Whether you are using gene expression, cloning and quantitative PCR, comparative genomic hybridization to detect copy number variation or selecting for the human exome to sequence regions of interest, Agilent provides the innovation and reliability your research requires.

Cytogenetics

Agilent’s portfolio of CGH (Comparative Genomic Hybridization), CGH+SNP (Comparative Genomic Hybridization plus Single Nucleotide Polymorphism), and CNV (Copy Number Variant) microarrays offers researchers access to a complete oligo array CGH solution for mapping large and small structural variations associated with developmental abnormalities, cancer development and progression, disease susceptibility, and differential drug responses.

Agilent Technologies Target Enrichment Solutions

Agilent’s SureSelect market leading platform provides a complete portfolio of catalog and custom products, providing the flexibility you need for your discovery to follow-up studies. New HaloPlex products combine the speed of PCR with the sensitivity of hybridization providing a robust solution for targeting smaller capture regions.

Bioinformatics Software - Integrate Data and Results from Multiple Applications

Agilent provides a full range of bioinformatics software tools to enable robust visualization and analysis of genomic datasets across a range of applications. GeneSpring software is the industry standard for gene expression analysis, miRNA and splicing analysis, and now includes DNA-Seq and RNA-Seq support for Agilent’s SureSelect target enrichment platform. Agilent’s bioinformatics solutions help researchers make the most of their data, and get answers they can trust.
ACM2

15 years of service to the customer

Delivery, installation and service on specialized equipment for analytical, scientific research, control and clinical laboratories

Thermo Scientific

- Chromatography
- Mass spectrometry
- elemental analysis
- Molecular spectroscopy
- Analysis on surfaces
- Software solution for laboratory information (LIMS)

ELGA

- Water treatment systems
- Ultra-pure water systems

1404 Sofia, ul. Tvardiski passage №23, et. 9, office 34
tel. 02 859 21 03; 02 859 21 30
fax: 02 958 28 18
e-mail: office@acm2.com
www.acm2.com
GS Junior System
The power of next-generation sequencing in your hands

Посетете ни на: www.chrono.bg

ХРОНО ООД, Вашият надежден партньор за Roche Applied Science в България.
Вашите партньори в сферата на високотехнологичните геномни решения

секвениране
метагеномика
родственост
транскриптотомика
геномика
епигенетика
генетика
генотипиране
генна експресия
фармакогеномика
онкология
генетика
трансляционна медицина
метагеномика
персонализирана медицина
генетичен скрининг