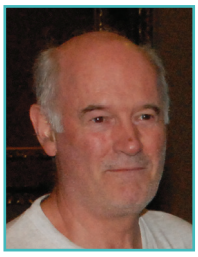
Oral Presentations

The exigencies of life at pH 1: evolution on earth and extra-terrestrial life signatures.



David S. Holmes§, Jorge Valdes, Francisco Duarte, Juan Pablo Cárdenas, Raquel Quatrini, Hector Osorio, Raúl Araya-Secchi, Tomás Pérez-Acle, Wendy González, Danilo González-Nilo, Eugenia Jedlicki

Center for Bioinformatics and Genome Biology, Facultad de Ciencias Biológicas, Universidad Andrés Bello, Chile,
§Speaker

Multiple comparative genome analyses and experimental evidence provide insight into the multiple challenges for microbial metabolism at pH 1. Acidophilic microorganisms are confronted by a delta pH across their membranes that are 6 orders of magnitude greater than that encountered by neutrophilic organisms. This DpH challenges the operation of membrane transporters that are driven by proton motive force (PMF). Models have been constructed by molecular simulation techniques that provide insight into how membrane transporters function in acid pH. A database has also been created for comparing physico-chemical properties of acidophilic and neutrophilc proteins. Another challenge confronting microorganisms at pH 1 is the concentration of soluble iron that can be 1016 times greater than that in neutrophilic environments. This poses a serious challenge to biological iron uptake and homeostatic mechanisms. However, the enormous delta pH also provides a unique opportunity to push electrons uphill against a thermodynamically unfavorable gradient to gain reducing power (NADPH). It also permits the microorganisms to synthesize ATP without the expenditure of conventional energy sources. Acid-driven, reverse PMF may have been an important energy source during the early evolution of life on Earth. In addition, the ability of these microorganisms to oxidize (and reduce) iron and sulfur, including solid substrates such as pyrite (FeS), has hallmarks of suggested primitive metabolic processes. These biological oxido-reduction reactions leave characteristic iron and sulfur isotope fractionation signatures and form minerals such as jarosite that are being used to search for evidence of ancient life on Mars. Also, the sulfuric acid rich ocean of Jupiter's moon Europa could be conducive to life forms with similarities to modern-day terrestrial acidophiles.

Regulation of gene expression: at the cross roads of novel high throughput studies, conceptual design and dynamical modeling.

Julio Collado-Vides

Center for Genomic Sciences, UNAM, Mexico

Research in our lab is in the middle of a transition at different levels, all of them associated with expansions in the field and in our vision of how we conceive RegulonDB, the database on transcription initiation and operon organization of *E. coli* K-12. We are working towards a RegulonDB conceived as a model of all genetic interactions, supported by classic experimental approaches –through the literature that we curate- as well as by novel high throughput (HT) technologies and their associated challenges. We also conceive it as a bioinformatic environment founded on knowledge of the network, predictive bioinformatic tools, and enabling dynamic predictions. We want to use this platform to generate an organized picture of the complete phenotypic repertoire –or **expressome**- of *E.coli*.

How do we address each of these changes? First of all, we need a new conceptual model of gene regulation with the precise terminology from signal, effector, core of regulation, adequate response, and eventual long-term ef-

fect. Sensing organs are proposed as the initial building blocks of a "reconstructive" or "synthetic" conceptual model of the cell. RegulonDB relational design derives from the conceptual model of the cell in terms of objects –molecules-, reactions and the regulation of their expression under different external conditions.

The HT mapping of promoters in the E.coli genome performed in the laboratory of Dr. Enrique Morett, is a collaboration that is posing many new challenges. A new perspective of transcriptional activity of the genome, more stochastic maybe, may well emerge. This will affect how we represent transcription at the level of promoters, individual genes, transcriptional units and operons. These HT datasets, the complete predictive bioinformatic "projections" of binding sites and promoters, as well as the genomic perspective of regulatory systems, is obligating us to re-evaluate classic notions of the regulation of gene expression in bacteria.

The biological challenge that all this knowledge and computational platform enable to address, is the one of understanding the structure of the phenotypic repertoire –the expressome or complete repertoire or total number of available configurations of expression that the *E.coli* genome enables. We have analyzed a large collection of Affimetrix microarrays with no less than 300 different conditions, and it has been shown that they group in around 70 different expression patterns. Understanding the contribution of different anatomical features of the network to reduce or "canalize" the expression space is an exiting challenge ahead.

Progress on the High-throughput Sequencing Projects at the Brazilian Bioinformatics Laboratory

Ana Tereza Vasconcelos[§], L.G.P. Almeida, M.E. Cantão, R.C. Souza, L.P. Ciapina, F.G. Barcellos, R.S. Silveira, M. Dellamano, A. Gerber, M.F. Nicolás

The National Laboratory for Scientific Computing, Brazil,
[§]Speaker

Bioinformatics Laboratory (Labinfo) is part of The National Laboratory for Scientific Computing LNCC that belongs to The Brazilian Ministry of Science and Technology (MCT). Labinfo has several projects dealing with genomics to systems biology with many genome sequencing under its responsibility (Brazilian Genome, South Genome, Nitrogen-Fixing Bacteria, Comparative *Xylella* spp., and *Burkholderia* spp.). The Labinfo team has developed the software SABIA (System for Automated Bacterial genome Integrated Annotation) [1] that assemblies, annotates and compares the genomes of prokaryotes, ESTs, eukaryotes and metagenomics. This team has also developed several databases on different themes that are available in its Website (http://www.labinfo.lncc.br).

In 2008, the LNCC constructed a new Lab for working together with the Labinfo, the Computational Genomics Laboratory (Unidade Genômica Computacional, UGC), which has three technicians and two researches as the current staff. Its mission is to sequence and analyze genomes, exomes, transcriptomes and metagenomes, using the Genome Sequencer System FLX Roche/454 Life Sciences and Bioinformatics tools. So far, the focus of our work is the genome sequencing (through several approaches) and analysis of a breast cancer cell line (HCC1954), various metagenomics samples, nitrogen-fixing bacteria (*Bradyrhizobium* spp. and *Rhizobium* spp.), an insect pathogenic fungus (*Metarhizium anisopliae*), a multicellular magnetotactic prokaryote from a hypersaline environment (*Candidatus Magnetoglobus multicellularis*), as well as the pandemic (Influenza A H1N1) 2009 virus (whole and partial). Up to now, through Genome Sequencer FLX Roche/454 Life Science we were able to perform 43 runs getting more than 35 millions of long reads (median 400 bp) that accounted almost 14 Gb. Currently we are applying the Newbler (provided by Roche) and the Celera, both assembly software systems and the SABIA for the annotation process.

References

1. ALMEIDA L. G. P., VASCONCELOS A. T. R. et al. A System for Automated Bacterial Integrated Annotation - SABIÁ. Bioinformatics (Oxford), England, v. 20, p. 2832-2833, 2004.

Studying RNA-binding proteins and their interactome in post-transcriptional networks

Sarath Chandra Janga[§], Nitish Mittal, Madan Babu M.

MRC Laboratory of Molecular Biology, University of Cambridge, UK,
[§]Speaker

Background

Messenger RNAs have traditionally been viewed as passive molecules in the pathway from transcription to translation. However, it is now clear that RNA-binding proteins (RBPs) play an important role in RNA metabolism by controlling gene expression at post-transcriptional level.

Methodology

The complete list of annotated RBPs and the data for RBP-RNA interactions in *Sacchoromyces cerevisiae* was obtained from Hogan et al (1). The total number of annotated RBPs in yeast reported in this study was 561 and mRNA targets for 41 RBPs have been systematically identified on a whole genome scale by employing the RIP-chip technology. A total of 14,312 interactions comprising of 41 RBPs and 5025 genes in the entire genome of *S. cerevisiae*, which forms a network of post-transcriptional interactions between RBPs and the target RNAs obtained using this technology, was used in this study (1). Protein interactions in *S. cerevisiae* were obtained from Collins et al (2).

Results

Here, we first show that RBPs as a functional class of proteins exhibit significantly higher proportion of protein-protein interactions compared to other proteins not encoding for RBPs (non-RBPs) in *S. cerevisiae*. We then extend our results to demonstrate that the number of RNA substrates bound by a RBP correlates positively with its number of protein interactions, using RBP-RNA networks for a selected set of RBPs. Functional analysis of the physically interacting proteins of RBPs suggests that RBPs can be grouped into distinct biological processes based on the gene ontology annotations of the physically interacting protein partners and on their degree, with splicing and translation-associated ones among highly interacting RBPs and mRNA transport associated with poorly connected RBPs. Further analysis to study post-translational modifications in RBPs using currently available kinase-substrate maps clearly revealed extensive role of phosphorylation. Our results reveal that RBPs are predominantly controlled at the protein level either through transient post-translational modifications or stable protein-protein interactions indicating their potential for regulation at the protein level to control diverse cellular processes.

Conclusions

These observations explain the molecular basis behind the cause of a number of disorders associated with aberrations in the expression levels of RBPs and provide a framework to elucidate the link between different levels of regulation in higher eukaryotes.

References

1. Hogan, D.J., et al. (2008) Diverse RNA-binding proteins interact with functionally related sets of RNAs, suggesting an extensive regulatory system. PLoS Biol 6, e2552.
2. Collins, S.R., et al. (2007) Toward a comprehensive atlas of the physical interactome of Saccharomyces cerevisiae. Mol Cell Proteomics 6, 439-450

Chipster – user-friendly software for reproducible analysis of high-throughput data


Aleksi Kallio, Jarno Tuimala, Taavi Hupponen[§], Petri Klemelä, Massimiliano Gentile, Eija Korpelainen[§]

CSC - IT Center for Science, Finland,
[§]Speaker

Background

High throughput technologies, such as microarrays and next generation sequencing provide enormous possibilities for genome wide studies of gene expression and regulation. The technological advances have been accompanied by active research in data analysis, producing new methods at rapid pace. While most of the newly developed methods are freely available, their use requires substantial computing skills, such as knowledge of the R programming language in the case of Bioconductor.

Method

In order to enable biologists with no programming background to benefit from the method development in a timely manner, we have developed the Chipster software (http://chipster.csc.fi/). Chipster brings a comprehensive collection of up-to-date analysis methods and visualization tools within the reach of bioscientists via its graphical user interface. Chipster has been mainly used for microarray data, but it is a generic software and we are currently implementing functionality for ChIP-seq and RNA-seq analysis.

Results

Chipster is a client-server system, where the client software utilizes Java Web Start technology for automatic installation and updates. The actual analysis modules, R libraries, annotations, and promoter and pathway databases are installed and updated centrally on the server side. The centralized approach reduces the maintenance burden and enables the analysis jobs to benefit from the CPU and memory of central computing servers. In fact Chipster's flexible architecture allows the computations to be distributed to several servers, and the tool collection can be further expanded by connecting external Web services to the system. In order to deal efficiently with large high-throughput datasets, we are currently investigating methods for data compression and transfer recovery. We have also developed a visualization framework that allows efficient random access and data sampling to produce interactive visualizations from large datasets.

Chipster enables users to save their analysis pipelines as reusable workflows, which can be shared with others. The system keeps track of the analysis steps taken and displays them visually in a workflow graph. Users can experiment with different methods and parameters, and prune the resulting workflow by simply deleting the unwanted steps. The ability to reuse and share workflows not only speeds up the analysis, but also improves consistency and provides an easy way for bioinformaticians to collaborate with biologists. Chipster is open source, and institutes can easily tailor it to their needs by adding new analysis tools using a simple mark-up language.

Conclusions

The open source software Chipster enables bioinformatics service providers to offer their users a biologist-friendly access to up-to-date analysis methods of high throughput data. Chipster facilitates reproducible and collaborative research by enabling users to save the performed analysis steps as reusable workflows, which can be also shared with others.

Next Generation Sequencing Methodologies: Applications in Comparative Genomics and Transcriptomics

Enrique Morett[§], Alfredo Mendoza, Leticia Olvera, Maricela Olvera, Ricardo Grande, Veronica Jimenez, Blanca Taboada, Leticia Vega, Katy Juarez, Heladia Salgado, Araceli Huerta, Julio Collado

Institute of Biotechnology, UNAM, Mexico,
[§]Speaker

Background

Recent advances in massively parallel high-throughput sequencing technologies have dramatically reduced the cost of nucleic acids sequencing by orders of magnitude and at the same time improved accuracy. Even when these technologies have been available by no more than three years, they have transformed our understanding of many aspects of biology and medicine, as genome variation in healthy and diseased individuals, including SNPs detection and DNA rearrangements; alternative splicing; epigenomics, as DNA methylation; transcriptomics and small RNA detection; and the interaction of proteins with DNA and RNA. Resequencing a human genome now can be done by a few thousand dollars in a matter of weeks. The great challenge now is to make sense of all the new data being generated.

Methods

We have used Next Generation Sequencing (NGS) technologies, in particular Roche's 454 pyrosequencing and Illumina's sequencing by synthesis platforms to understand transcription regulation at a genomic scale in *E. coli* and in Geobacter sulfurreducens.

Results and Conclusions

By determining transcription start sites and therefore regulatory regions, more than 2000 new promoters in each one of these organisms have been found. Interestingly, we have detected multiple promoters within operons and in antisense orientation. We propose that many of these antisense transcripts play a regulatory role still poorly understood. We have developed dedicated bioinformatics tools to help in data analysis. With this wealth of data we are now able to study gene regulation at a global scale.

Systems Biology for PCR detection of malaria at the liver stage

Ezekiel Adebiyi[§], Victor Osamor, Seydou Doumbia

Department of Computer and Information Sciences, Covenant University, Nigeria,
[§]Speaker

The major aim of this work is to develop an efficient and effective k-means algorithm to cluster malaria microarray data to enable the extraction of a functional relationship of genes for malaria treatment discovery. However, traditional k-means and most k-means variants are still computationally expensive for large datasets such as microarray data, which have large datasets with a large dimension size d. Huge data is generated and biologists have the challenge of extracting useful information from volumes of microarray data. Firstly, in this work, we develop a novel k-means algorithm, which is simple but more efficient than the traditional k-means and the recent enhanced k-means. Using our method, the new k-means algorithm is able to save significant computation time at each iteration and thus arrive at an O(nk2) expected run time. Our new algorithm is based on the recently established relationship between principal component analysis and the k-means clustering. We further prove that our algorithm is correct theoretically. Results obtained from testing the algorithm on three biological data and three non-biological data also indicate that our algorithm is empirically faster than other known

k-means algorithms. We assessed the quality of our algorithm clusters against the clusters of known structure using the Hubert-Arabie Adjusted Rand index (ARIHA), we found that when k is close to d, the quality is good (ARIHA > 0.8) and when k is not close to d, the quality of our new k-means algorithm is excellent (ARIHA > 0.9). We compare three different k-means algorithms including our novel Metric Matrics k-means (MMk-means), results from an in-vitro microarray data with the classification from an in-vivo microarray data in order to perform a comparative functional classification of P. falciparum genes and further validate the effectiveness of our MMk-means algorithm. Results from this study indicate that the resulting distribution of the comparison of the three algorithms' in- vitro clusters against the in-vivo clusters is similar, thereb authenticating our MMk-means method and its effectiveness. Lastly using clustering, R programming (with Wilcoxon statistical test on this platform) and the new microarray data of *P. yoelli* at the liver stage and the *P. falciparum* microarray data at the blood stages, we extracted twenty nine (29) viable *P. falciparum* and *P. yoelli* genes that can be used for designing a Polymerase Chain Reaction (PCR) primer experiment for the detection of malaria at the liver stage. Due to the intellectual property right, we are unable to list these genes here.

Phylogenetic profiling in parallel with high through put screens implicate new genes in the RNAi pathway

Yuval Tabach[§], John Kim, Harrison W. Gabel, Ravi S. Kamath, Gary Ruvkun
Massachusetts General Hospital, Harvard Medical School, US,
[§]Speaker

Small RNAs are short length single-stranded RNA molecules, which regulate genes and genomes. This regulation can occur at some of the most important levels of genome function, including chromatin structure, chromosome segregation, transcription, RNA processing, RNA stability, and translation. Since the first discovery of miRNA in 1993 and subsequent uncovering of numerous classes of small RNA molecules, the molecular biology of small RNA has become significant hub in biological and medical research.

While related small RNA pathways are conserved across diverse phylogeny they have been lost in some specific species such as the budding yeast Saccharomyces cerevisiae. Proteins with specific roles in small RNA pathways show highest conservation in organisms that maintain these pathways, while they are lost, or diverged in organisms without them. For example the Argonaute family of proteins which are central to small RNA function are highly conserved in many organisms with the RNAi pathway but lost in budding yeast. The pattern of conservation and loss of small RNA pathways across the eukaryote phylogenetic tree makes it ideal for phylogenetic profile analysis. A phylogenetic profile describes the presence or absence of a protein in a set of reference genomes. To comprehensively analyze the predictive value of phylogenetic profiling in analysis of specific biological pathways, we have generated phylogenetic profiles for the entire worm proteome by analyzing the similarity of all the ~20,000 predicted worm proteins across all available eukaryote genomes. We have analyzed the phylogenetic similarity profile of genes in small RNA pathways with the hypothesis that the greater the phylogenetic profile similarity is between genes, the greater the likelihood of proteins sharing membership in the same pathway or cellular system. We clustered the profiles and merged it with genome-wide RNAi screens of microRNA and RNAi pathway genes, as well as protein interaction data generated by yeast two hybrid and immune precipitation analyses.

While the overlap between the genes from these different datasets was small, these genes were significantly over-represented in several phylogenetic profiling clusters. These results suggest that these genes do in fact share biological functionality. One significant cluster contains most of the Argonaute and PIWI proteins in addition to several genes that previously were unknown to be involved in the RNAi process.

In conclusion we have run phylogenetic profiling analysis of the worm proteome on all finished sequenced eukaryotes. Our results suggest that this approach will have significant utility in understanding and merging results from different screens of related biological processes that may show low overlap. Finally, this analysis has implicated new genes that are closely linked to the Argonaute proteins in evolution, and like Argoanutes, likely function in small RNA pathways.

Central Core DNA Sequence Information System (CCSIS)

Alex Patak[§,1], Peter Henriksson, Alessia Maineri, Paolo Struffi, Guy Van den Eede

Molecular Biology and Genomics Unit, Institute for Health and Consumer Protection, Joint Research Centre, European Commission, Italy, [§]Speaker, [1]Node Manager

Background

The **Institute for Health & Consumer Protection** (IHCP) is one of the seven scientific institutes of the Joint Research Centre (JRC) of the European Commission.

The Molecular Biology and Genomics (MBG) Unit focuses on the DNA sequences of genetically modified crops (GMOs) and the methods derived thereof to detect, identify and quantify GMOs to comply with the EU regulations.

The MBG Unit hosts the "Community Reference Laboratory for GM Food and Feed" (CRL-GMFF) for the validation of GMO detection methods as part of the EU authorization procedure.

This work is coordinated at the European level in collaboration with the "European Network of GMO Laboratories" (ENGL) network.

The IHCP - EMBnet specialized node focuses on the "**Central Core DNA Sequences Information System**" (CCSIS) which is the molecular database where the submitted Genetically Modified Organism sequence data to the "Community Reference Laboratory for GM Food and Feed" (CRL) by applicants is stored to run **homology searches** in order to assess the specificity of the proposed GMO detection method as required by the COMMISSION REGULATION (EC) No 641/2004.

Method

All sequences provided by applicant companies or in-house sequenced GMO events, have been manually annotated and made available via web, with restricted user-access, on a dedicated computing platform integrated with bioinformatics tools.

The CCSIS can be divided into:

Hardware

- Platform based on an Apple Workgroup Cluster for Bioinformatics. The CCSIS consists of 4 Apple Xserve G5 and Xserve RAID (4.09 TB).

Software

- Bioteam iNquiry with over 200 bioinformatics applications and running on a Sun Grid Engine based cluster.
- MRS (M. L. Hekkelman) a search engine for biological and medical databanks.
- The NCBI's freely available standalone annotation and submission program Sequin has been used for the annotation.

Data

- All GMO sequences have been manually annotated following international annotation rules (International Nucleotide Sequence Database Collaboration's (INSDC DDBJ/EMBL/GenBank) Feature Table). The data covers GMO events, GMO Detection Methods, plasmid and reference genes.
- Local copies of public available databases are installed (Genbank, patents, plant genomes, etc).

Results

The CCSIS is now routinely in use at the CRL-GMFF for the scientific assessment of GMO detection methods and contains a unique and confidential GMO sequence dataset not publically available.

So far a total of 197 sequence records are available on CCSIS and distributed as following: 69 Single Events, 31 Crossings (Stacks), 62 PCR Amplicons of CRL Validated Detection Methods, 26 Reference Genes, 6 Plasmids and 3 In-House Sequenced.

The data is continuously updated and extended with new information.

Conclusion

CCSIS has demonstrated to be a valuable tool for the specificity assessment of GMO detection methods (CRL-GMFF) and also for the development of new GMO screening methods (MBG Unit).

Dynamical behavior and functional mechanism in biological Networks.

Osbaldo Resendis-Antonio[§], Pamela Silver

Center for Genomic Sciences, UNAM, Mexico, [§]Speaker

Background

With the advent of high throughput technologies, an immediate challenge has been the development of analytical procedures for integrating these data and contribute to grasp the biological mechanisms underlying (dys)functional phenotypes in cells. To this purpose, mathematical modeling of genetic and metabolic circuits has served as a guide to establish hypothesis that can be subsequently assessed at an experimental level. Even though there have been important achievements between models and its experimental counterpart in synthetic and systems biology, the lack of kinetic information is a bottleneck when one desires to analyze temporal behavior on some of these biological circuits. In this work, we present a methodology that involves high throughput technology for estimating dynamical properties of a perturbed metabolic network whose kinetic information associated to its set of reactions is lacking or incomplete.

Methods

The method involves two sources of information:

1. The stoichiometric matrix (integrating the biochemical reactions in an organism), whose reconstruction is based on accurate data bases, and
2. Metabolome data (characterizing a physiological steady state)

Taking into account both sources of information, and assuming that law mass action is valid for each reaction, we construct a library of dynamical systems whose statistical properties allows us to identify how fast and in what extend the metabolites temporally organize to reach its steady state after a perturbation. The library of dynamical system is constructed by two steps. Firstly, one proceed to identify and select some kinetic parameters that constraints the metabolic network at a steady state behavior on metabolites and fluxes. Consequently, the time scales emerged from classical linear theory perturbation were obtained from these set of parameters and its global statistical properties describing how and how fast the cell reach its steady state were analyzed.

Results

For the sake of simplicity, we illustrate the method in some biological networks whose dysfunction state is related with human disease. The previous analyses lead us to identify robustness along time scales. Conversely, one observe that the way by which the metabolites orchestrates the response towards equilibrium can be accomplished by a wide variety of mechanisms determined by specific kinetics parameters. Even though this broad variability, one can identify a handful of metabolites that invariantly participate along metabolic relaxation.

Conclusions

In this talk, we present a formalism to explore the feasiblespace of dynamical behavior emerged from a metabolic network. The overall scope of this framework is focused to estimate the magnitude of kinetic parameters in close relation with high throughput technology. Overall, this method represents a novel procedure to explore the potential relation between proper functionality and its characterization through dynamic behavior. Given the underlying assumptions, we expect that surveying the complex relationship between functionality and dynamics in metabolisms will be useful to identify dynamic variables that may be used to characterize (dys)functional metabolic states at a cell population level, a desired aim

toward the optimal design of drugs and again human diseases.

Gene relations (correlations and interactions) include critical information to build accurate disease-classifiers derived from gene expression profiles and to find disease-networks

Celia Fontanillo, Alberto Risueño, Javier De Las Rivas[§]

Cancer Research Center (CiC-IBMCC, CSIC/USAL), Spain,
[§]Speaker

In recent years, in the biomedical field several Machine Learning (ML) methods (such us k-NN, neural networks, random forest or SVM) have been applied to global gene expression data from genome-wide microarrays to build disease classifiers and predictors, which allow to identify different types and sub-types of diseases. These computational methods explore and compare the gene expression profiles across individuals (i.e. multiple patient clinical samples) in order to find the genes that best define each pathological state studied. Despite the broad interest to build disease classifiers, ML algorithms usually include some technical characteristics that can be problematic when they are applied to discover the biomolecular entities that are behind a biological state studied: (i) many ML algorithms make transformations that are not transparent to the features inside the classifiers, i.e. they do not allow to identify the genes that correspond to the key features which provide best discriminatory power; (ii) many ML algorithms use feature selection strategies to reduce feature redundancy, and it is not clear how some genes in a living organism can be considered "redundant" and "eliminated" without the risk of losing biological functions and meaning inside a given biological state that wants to be classified. Following these considerations, we chose for this study SVM (support vector machine) method with a linear kernel, because it is a type of generalized classifier which provides transparent direct correlation of the support vectors with the variable features (i.e. the genes) in the classification hyperplane. Adequate classifiers may reflect the biology behind the explored states, but, in principle, there is not a clear link between their "predictive power" and their "explanatory potential". Therefore, the second problem described brought us to investigate how common strategies applied by ML methods to eliminate or reduce feature redundancy affect disease classifiers derived from microarrays gene expression data, and if such strategies leaded to gain or loss of biological meaning.

Sequencing, assembly and comparative genomics of several Staphylococcus aureus Straits

Laurent Falquet[§], Sandra Calderon, Valérie Vogel, Patrick Basset, Dominique Blanc

Swiss Institute of Bioinformatics, Switzerland,
[§]Speaker

Staphylococcus aureus is a well known opportunistic bacteria often found in hospital acquired infections. Understanding its microevolution should highlight the mechanism by which genetic factors might be transmitted, resulting in the emergence of new clones with specific biological characteristics, such as pathogenicity, virulence, or antibiotic resistance. Comparative genomics of different strains should allow the identification of the genetic basis of such characteristics. When the strains are closely related, the direct comparison of the contigs obtained by de novo assembly might be unsatisfactory and the comparison with a reference genome could be required. However assembling by mapping onto

a reference genome often only identifies the similarities, one must carefully analyse the non-matched orphaned reads in order to identify the real differences.We present preliminary results of Ultra High Throughput Illumina sequencing and comparison by mapping onto a reference genome, as well as de novo assembly of non-mapped reads.

Next Generation Sequencing techniques and Web 2 technologies

Erik Bongcam-Rudloff

Chairman, European Molecular Biology Network (EMBnet). Swedish University of Agricultural Sciences (SLU), Uppsala, Sweden.

With the recent introduction of Next Generation (NextGen) sequences technologies to the Life Sciences arena, the volume of sequence data being created is growing at an astonishing rate. Furthermore, high-throughput genomics and proteomics technologies are becoming more common worldwide and these technologies are evolving at a rapid pace. Sequencing platforms facilities are being established in research institutions in all continents making the new techniques accessible to researchers in a broad range of fields.

Questions can be posed in new and different ways with more large-scale methods. For instance, NextGen technologies makes it possible to map the bacteria flora in a person's mouth, to see why one individual develops infectious diseases like e.g. malaria while another does not. The new techniques also make it possible to monitor how the vectors and host adapts (mosquitoes, parasites, virus) in order to escape to people's immune defenses.

In my talk I will address several aspects related to the handling of the avalanche of biological data and list ways of extracting information from these data. Sharing research data is essential for effective collaboration, I will therefore put special emphasis in discussing how we can make use of Web 3.0 technologies in the Life Sciences areas.

To produce sequence data today is easy and cheap, to analyze and annotate is the difficult, tedious and time consuming part. In the near future we will need armadas of biologists with basic bioinformatics skills, this open opportunities for the bioinformatics communities outside the traditional bastions. I will discuss those aspects and present new ideas.