XI Jornadas de Bioinformática

January 23-25th, 2012 @ PRBB/BARCELONA http://jbi2012.org

BOOK OF ABSTRACTS

OUR THANKS TO:

INB Spanish National Bioinformatics Institute

Generalitat de Catalunya
Departament d'Economia
i Coneixement

PORTUGUESE BIOINFORMATICS

biobam

INBIOMEDvision

Genoma España

ORGANIZATION SUPPORT:

cnag

UNIVERSITAT POMPEU FABRA

ISCB Student COUNCIL

prbb Barcelona Biomedical Research Park

Welcome to JBI2012!

The XIth Spanish Symposium on Bioinformatics (JBI2012), which takes place in Barcelona between the 23rd and 25th of January 2012, is co-organised by the Spanish Institut of Bioinformatics and the Portuguese Bioinformatics Network and hosted by the Barcelona Biomedical Research Park (PRBB).

We welcome your all to this year's conference that aims at bringing together the Bioinformatics community of the South of Europe and North of Africa. This year, the reference topic is "*Genome Architecture, Annotation and Design*" for which the conference will provide the opportunity to discuss the state of the art for the integration of the fields of biology, medicine and informatics.

We really hope that you will enjoy the meeting!

MARC A. MARTI-RENOM

Baldomero Oliva (IMIM-UPF)

Marc A. Marti-Renom (CNAG)

On behalf of the JBI2012 **Organizing Committee:**

Arcadi Navarro (UPF, Barcelona)
Ben Lehner (CRG, Barcelona)
Claudio Soares (ITQB, Lisboa)
Ferran Sanz (UPF, Barcelona)
Francesc X. Avilés (UAB, Bellaterra)
Modesto Orozco (IRB, Barcelona)
Patrick Aloy (IRB, Barcelona)
Roderic Guigó (CRG, Barcelona)
Xavier Daura (UAB, Bellaterra)

the Student Symposium **Chairs**:

Lorena Pantano (IMPPC, Badalona)
Salvador Capella (CRG, Barcelona)

and the JBI2012 **Secretariat**:

Martina Gasull Masip (IMIM-UPF, Barcelona)
Carina Oliver Dutrem (IMIM-UPF, Barcelona)

Student Symposium Program

Monday 23rd, 2012

9:00AM	**Student Symposium Registration**	**Student Symposium Poster Session**
10:00AM	**Student Symposium Keynote (Chair: Salvador Capella)** • **Ben Lehner (CRG, Barcelona)** *Predicting the phenotypes of individuals: why would a mutation kill me, but not you?*	
11:00AM	Coffee break	**Student Symposium Poster Session**
11:30AM	**Student Symposium Session 1 (Chair: Lorena Pantano)** • **Eneritz Agirre (UPF, Barcelona)** *The role of epigenetics and small RNAs in the regulation of pre-mRNA splicing.* • **Inna Povolotskaya (CRG, Barcelona)** *Stop codons in bacteria are not selectively equivalent.* • **João Curado (CRG, Barcelona & GABBA/UP, Porto)** *Deep sequencing of RNA from distinct sub-cellular fractions shows that splicing in the human genome occurs predominantly during transcription.*	
1:00PM	Lunch	**Student Symposium Poster Session**
2:15AM	**Student Symposium Session 2 (Chair: Ana Fernandes Oliveira)** • **Núria Radó-Trilla (FIMIM, Barcelona)** *Low-complexity regions as a mechanism of protein diversification.* • **Yassine Souilmi (MVU, Rabat)** *SNP4Forensic Project.* • **Hiren Karathia (UdLL, Lleida)** *Homol-MetReS: A web application for integration between molecular systems biology and evolutionary biology.*	
3:45PM	**JBI2012 Welcome**	
4:00PM	**JBI2012 Keynote #1.**	
5:00PM	**Student Symposium "Job Offers" Session**	**INB Meeting**
8:00PM	**Student Symposium Social Event**	

JBI2012 Program

Monday 23rd, 2012

2:00AM to 3:30PM	**JBI2012 Registration**
3:45PM	**JBI2012 Welcome**
4:00PM	**JBI2012 Keynote #1. Chair: Ferran Sanz** • **Søren Brunak (CBS, Denmark)** *Interfacing sequencing and network biology data to personal healthcare sector information.*
5:00PM	**INB Session. Chair: Christian Blaschke** ***Current challenges and limitations for bioinformatics researchers.*** *In this session a group of expert bioinformaticians will discuss their views on where the current challenges and limitations lie that researchers in this field experience. In continuation the attendees of the session will have the possibility to address questions to the invited experts and offer their own take on the subject. This will be a very interactive and lively session that we hope will serve both newcomers and veterans in bioinformatics to enrich their views around these interesting questions.*

Tuesday 24th, 2012

9:00AM	**JBI2012 Keynote #2 (Chair: Marc A. Marti-Renom)** • **Leonid Mirny (MIT, Boston)** *Higher-order chromatin architecture: bridging physics and biology.*	
10:00AM	Coffee break	**Poster Session I (odd posters)**
10:30AM	**JBI2012 Session 1. Genome Architecture (Chairs: Eduardo Eyras & Ivo Gut)** • **Simon C. Heath (CNAG, Barcelona)** *Joint estimation of methylation probability and genotype from whole genome bisulfite sequence data.* • **Juan Ramón Gonzalez (CREAL, Cerdanyola)** *Gene-specific count data distributions are required in RNA-seq experiments with extensive replication.* • **Eva Maria Novoa (IRB, Barcelona)** *A Role for tRNA Modifications in Genome Structure and Codon Usage.* • **Ignacio Medina (CIPF, Valencia)** *The first HPC pipeline for Next Generation Sequencing data analysis.* • **Tomas Marques-Bonet (ICREA-IBE, Barcelona)** *Structural variation from next generation sequencing. Limits of the technology and lessons from the Great Ape Genome Project.*	
12:30PM	Lunch	**Poster Session I (odd posters)**
2:30PM	**JBI2012 Session 2. Highlights 2011 (I) (Chairs: Modesto Orozco & Alfonso Valencia)** • **Davide Baù (CIPF, Valencia)** *The three-dimensional folding of the α-globin gene domain reveals formation of chromatin globules.* • **Tanya Vavouri (CRG, Barcelona)** *Chromatin organization in sperm may be the major functional consequence of base composition variation in the human genome.* • **Javier Macia (UPF, Barcelona)** *Distributed biological computation with multicellular engineered networks.*	
4:00PM	Coffee break	**Poster Session I (odd posters)**
4:30PM	**JBI2012 Session 3. Structural Bioinformatics (Chairs: Patrick Aloy & Cláudio Soares)** • **Toni Giorgino (GRIB-IMIM, Barcelona)** *Molecular recognition of SH2-phosphopeptide by molecular dynamics.* • **Pablo Minguez (EMBL, Heidelberg)** *A global network of crosstalking post-translational modifications.* • **Joan Planas-Iglesias (GRIB-IMIM, Barcelona)** *To Bind or not To Bind: Predicting protein-protein interactions from favouring and disfavouring local structural features.* • **Antonio Morreale (CBM-CSIC, Madrid)** *A reverse combination of structure-based and ligand-based strategies for virtual screening.*	
8:00PM	**Gala dinner**	

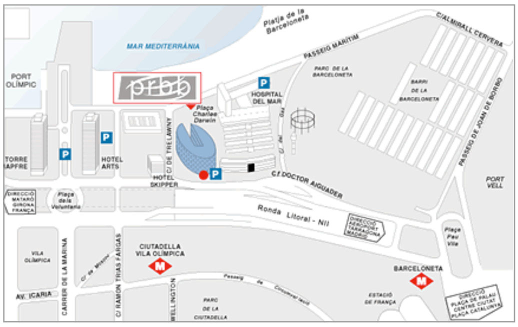
Wednesday 25th, 2012

9:00AM	**JBI2012 Keynote #3 (Chair: Baldomero Oliva)** • **Luís Serrano (CRG, Barcelona)** *A quantitative systems biology study on a model bacterium.*	
10:00AM	Coffee break	**Poster Session II (even posters)**
10:30AM	**JBI2012 Session 4. Genome Annotation (Chairs: Ana Teresa Freitas & Roderic Guigó)** • **Antonio M. Mérida (UM, Málaga)** *Sma3s: a 3 stages software for sequences make sense.* • **Paolo Ribeca (CNAG, Barcelona)** *The GEM toolkit: world-class short read mapping, 100% made-in-Spain.* • **M. Gonzalo Claros (UM, Málaga)** *Highly efficient pre-processing of NGS reads and identification of full-length genes.* • **Beatriz García-Jiménez (UCIII, Madrid)** *Relational Learning-based Extension for Reactome Pathways with Sequence Features and Interactions.*	
12:30PM	Lunch	**Poster Session II (even posters)**
2:30PM	**JBI2012 Session 5. Highlights 2011 (II) (Chairs: Modesto Orozco & Alfonso Valencia)** • **Patrick Aloy (IRB, Barcelona)** *Interactome mapping suggests new mechanistic details underlying Alzheimer's disease.* • **Ana M. Rojas (IMMPC, Badalona)** *The RAS Superfamily of signaling proteins: a 2011 update.* • **Mar Gonzàlez-Porta (EBI, Cambridge)** *Estimation of alternative splicing variability in human populations.*	
4:00PM	Coffee break	**Poster Session II (even posters)**
4:30PM	**JBI2012 Session 6. Phylogenetics and evolution (Chairs: José Pereira Leal & Arcadi Navarro)** • **Alberto Pascual-García (CBM-CSIC, Madrid)** *Detecting bacterial interactions from environmental samples: Ecological aggregations favor bacterial cosmopolitanism.* • **Hernán Dopazo (CIPF, Valencia)** *Does Nature Play Dice with Genomes?* • **Jaime Huerta-Cepas (CRG, Barcelona)** *Nested Phylogenetic Reconstruction: scalable resolution in large phylogenies.* • **Urko M. Marigorta (IBE-CSIC, Barcelona)** *Recent human evolution, continental differences in genes for complex disease and the common gene/common variant hypothesis.* • **Antonio Barbadilla (IBB-UAB, Cerdanyola)** *Population genomics of 158 genomes of* Drosophila melanogaster.	
6:00PM	**Communication Awards and Concluding Remarks**	

Maps & Location

The Student Symposium and the JBI2012 meetings take place at the PRBB in Barcelona.

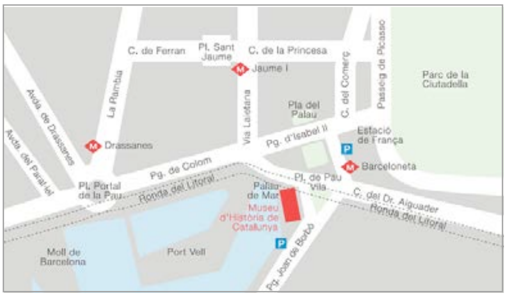
Parc de Recerca Biomèdica de Barcelona (PRBB)
c/ Dr. Aiguader, 88
E-08003 Barcelona
Tel. (+34) 93 316 0000
Fax (+34) 93 316 0019
E-mail: comunicacio@prbb.org
Web: www.prbb.org



- Student Symposium: **Eivissa Room** (PRBB, 1st floor)
- Student Symposium Poster Session: **Ramon y Cajal Room** (ground floor)
- JBI2012 talks/presentations: **PRBB Auditori**
- Poster sessions and lunch/coffee breaks **underneath** of the PRBB Auditorium.

The Gala dinner on the 24th of January takes place at the covered rooftop terrace of the "Museu d'Història de Catalunya", which is located only few blocks south of the PRBB.

Museu d'Història de Catalunya (MHCAT)
Pl. de Pau Vila, 3 (Palau de Mar)
08003 BARCELONA (Barcelonès)
Tel: (+34) 93 225 47 00
E-mail: mhc.cultura@gencat.net
Web: http://www.en.mhcat.net



Student Symposium Keynote

Monday 23rd, 2012 @ 9:00AM

Ben Lehner
Group Leader
EMBL-CRG Systems Biology Research Unit and
ICREA, Centre for Genomic Regulation, UPF, Barcelona, Spain

Ben Lehner studied Natural Sciences at the University of Cambridge, followed by a PhD in genomics (human protein interaction networks, antisense transcription). He was then a post-doc in the Fraser lab at the Wellcome Trust Sanger Institute, where he started working with *C. elegans* and on genetic interactions. Since December 2006 he has been a group leader in the EMBL-CRG Systems Biology Program at the CRG, and since 2009 an ICREA Research Professor. Recent work has focussed on the biology (genetics, epigenetics) of individuals, genome organization, and on the development of new methods. The lab is funded by the ERC, EMBO Young Investigator Program, MICINN, AGAUR, ERASysBio+ and ICREA..

Predicting the phenotypes of individuals: why would a mutation kill me, but not you?

To what extent is it possible to predict the phenotypic differences among individuals from their completely sequenced genomes? We use model organisms (yeast, worms) to understand when you can, and when you cannot, predict the biology of an individual from their genome sequence.

Keynote #1

Monday 23rd, 2012 @ 4:00PM

Søren Brunak
Professor
Technical University of Denmark. Lyngby, Denmark.
Founding director of the Center for Biological Sequence Analysis

Søren Brunak, Ph.D., is professor of Bioinformatics at the Technical University of Denmark and professor of Disease Systems Biology at the University of Copenhagen. Prof. Brunak is the founding Director of the Center for Biological Sequence Analysis, which was formed in 1993 as a multidisciplinary research group of molecular biologists, biochemists, medical doctors, physicists, and computer scientists. Søren Brunak has been highly active within data integration, where machine learning techniques often have been used to integrate predicted or experimentally established functional genome and proteome annotation. His current research does combine molecular level systems biology and healthcare sector data such as electronic patient records and biobank questionnaires . The aim is to group and stratify patients not only from their genotype, but also phenotypically based on the clinical descriptions in the medical records.

Interfacing sequencing and network biology data to personal healthcare sector information.

World-wide the healthcare sector is confronted with the availability of database information which describe the individual in great detail. These data range all the way from the molecular level, where they for example reveal the genetic makup of the patient, to the fine-grained descriptions of disease phenotypes as they are found in electronic patient records at hospitals. Linking these data is a huge undertaking which soon will represent a major challenge given that it already has become feasible to sequence the DNA of entire populations at low cost. Combining molecular level data with clinical information and data on the chemical environment may add complementary types of knowledge which – together with genotype and metagenomic information from the individual – can reveal disease mechanisms in novel ways. Electronic patient records remain a rather unexplored, but potentially rich data source for example for discovering correlations between diseases. We describe a general approach for gathering phenotypic descriptions of patients from medical records in a systematic and non-cohort dependent manner. By extracting phenotype information from the free-text in such records we demonstrate that we can extend the information contained in the structured record data, and use it for producing fine-grained patient stratification and disease co-occurrence statistics. The approach uses a dictionary based on the WHO International Classification of Disease ontology and is therefore in principle language independent.

Reference: *Using electronic patient records to discover disease correlations and stratify patient cohorts.* Roque FS, Jensen PB, Schmock H, Dalgaard M, Andreatta M, Hansen T, Søeby K, Bredkjær S, Juul A, Werge T, Jensen LJ, and Brunak S. PLoS Comput Biol. 2011 Aug;7(8):e1002141.

Keynote #2

Tuesday 24th, 2012 @ 9:00AM

Leonid Mirny
Associate Professor
Division of Health Sciences and Technology and
Department of Physics, MIT. Cambridge (MA), USA

Leonid Mirny got his masters degree in Chemistry at the Weizmann Institute in Israel and Ph.D. in Biophysics at Harvard, where he worked on theoretical protein folding. Since 2001 he has been a faculty member at MIT he worked on a range of problem in genomics and biophysics and most recently on 3D structure of the genome.

Higher-order chromatin architecture: bridging physics and biology.

Biophysical studies of chromatin provide new biological insights and pose new challenges for physics. Human genomic DNA is folded into a structure that fits in cell nucleus, while allowing chromatin reorganization required by many cellular processes. Recently developed Chromosome Conformation Capture method, Hi-C, provides a comprehensive and unbiased way of mapping interactions between all genomic loci. Our molecular simulations and analysis of Hi-C data for human cells suggest that on the Mb scale the chromatin fiber is folded into a fractal (crumpled) globule, a non-equilibrium state of a polymer. The fractal globule structure is ideally suited for chromatin organization during interphase: the fiber is densely packed, while forming no knots thus allowing easy unfolding and folding of sub-domains, e.g. during gene activation and repression. We use simulations to test mechanisms responsible for formation of the fractal globule during the cell cycle and formation of distinct domains of active and inactive chromatin. Furthermore by analyzing data on chromosomal abnormalities in cancer we establish a remarkable connection between them and higher-order chromatin architecture: spatial contacts between loci increase the probability of chromosomal exchanges between them. Comparison of Hi-C data for human and yeast cells suggest that the same physical principles lead to formation of different chromosomal architectures in these organisms.

Keynote #3

Wednesday 25th, 2012 @ 9:00AM

Luís Serrano

CRG Director, EMBO member and ICREA professor. Barcelona, Spain.

Luis Serrano did his PhD at the CBM (Madrid, Spain) on Cell Biology. Then he spent 4 years in the laboratory of Prof. A.R. Fehrs (MRC, UK) working in protein folding. In 1993, he became Group Leader at the EMBL (Heidelberg, Germany) working in Protein Folding and design. Ten years later, he was appointed head of the Structural & Computational Biology programme at the EMBL and he started to work on Systems Biology. By the end of 2006 he moved back to Spain to lead a programme working on Systems Biology, where he was appointed vice-director before finally becoming the CRG director last July 2011. His group is currently focused on Synthetic Biology, engineering and designing of biological systems. He is EMBO and RACEFyN member and received the Marie Curie Excellence Award. He has also been awarded with the prestigious ERC Advanced Grant and participates as Principal Investigators in many research projects financed both by the European Commission (through the 6th and 7th Framework Programmes) and the Spanish Ministry of Science and Innovation. He is Professor of ICREA and has directed 12 PhD thesis. He has published more than 240 papers in international journals. He has always been very mindful about the importance of the successful transfer of scientific discoveries to the society. He was involved in the creation of one of the first Spanish Biotech. Companies (Diverdrugs) in 1999. He is also co-founder of Cellzome, EnVivo and TRISKEL. Some of his work has been commented in Spanish newspapers (El Pais, LaVanguardia, El Mundo...), in the radio and other journals like Newsweek.

A quantitative systems biology study on a model bacterium.

The goal of Systems Biology is to provide a quantitative and predictive description of a living system to the extent that it can be fully simulated in a computer. We have undertaken such *Endeavour* using as a model the small bacterium *Mycoplasma pneumoniae*, a human pathogenic bacterium causing atypical pneumonia. Containing a reduced genome with only 690 ORFs, this bacterium is an ideal organism for exhaustive quantitative and systems-wide studies, avoiding technical limitations due to exceeding sample complexity, constrained by limitations in dynamic range and resolution of current generation mass spectrometers. Available data on the transcriptome, on protein complexes, as well as on metabolic pathways facilitate the integration of the data generated for this study into an organism-wide context. Additionally, *M. pneumoniae* represents a relevant organism to study stochastic noise in living systems. The cells are significantly smaller than other bacteria, such as *Eschericha coli* (0.05 mm^3 and 1 mm^3, respectively) resulting in principle in an increased susceptibility to abundance fluctuations of cellular molecules. Our analysis shows that even apparently simpler organisms have a large hidden layer of complexity and that for every question we have answered we have got two new ones. We are still far away to be get a full understanding of a cell.

Oral presentations

Session 1. **Student Symposium**
Monday 23rd @ 11:30AM

Student Symposium Session 1. Talk #1

The role of epigenetics and small RNAs in the regulation of pre- mRNA splicing.

Eneritz Agirre[1], Amadís Pagès[1], and Eduardo Eyras[1,2]

1.Computational Genomics Group, Universitat Pompeu Fabra, Dr. Aiguader 88, E08003 Barcelona, Spain. 2. Institució Catalana de Recerca i Estudis Avançats (ICREA), Passeig Lluís Companys 23, E08010, Barcelona, Spain

Every step in the control of gene expression undergoes dramatic alterations during cancer development, which can affect genes involved in proliferation, apoptosis, metabolism and invasion, thereby contributing to tumorigenesis. In particular, changes in the splicing regulation and in the chromatin state of genes have been both linked to the origin of cell transformation in many tumors. Interestingly, recent experiments have linked these two regulatory programs through the action of small non-coding RNAs [1]. Moreover, it has been shown that a direct interaction between chromatin and splicing factors is possible through adaptor proteins [2]. These facts indicates that there exists a complex regulatory network involving chromatin and RNA, which could be determined by a histone code, possibly under the control of small RNAs. This code may be essential for the normal function of the cell and its disruption may play a crucial role in cancer development.

Recent advances in high-throughput sequencing technologies provide a very effective way to obtain information about the regulation of genes by chromatin changes at a genome-wide level. These advances offer an opportunity without precedent to explore how splicing is regulated at a global level by epigenetic changes and by the effect of small RNAs. Probabilistic models that integrate all these datasets can help obtaining genome signatures of gene regulation and generate biological predictions. Recent publications have described probabilistic models relating epigenetic features to expression regulation [3], or a variety of sequence features to splicing regulation [4]. However, these methods have not yet explored the relationships between small RNAs, epigenetics and the splicing of pre-mRNAs.

Using a combination of measurable features for epigenetic and small RNA activity in genes, obtained by a systematic analysis of deep-sequencing data [5], we have applied various Machine Learning methods to classify Alternative Splicing events. We show that these features can be predictive of the splicing change for up to 80% of the cases. Moreover, quantifying the relevance of the data attributes by using feature selection methods, we obtain that the best descriptive features include a combination of epigenetic and small RNA features in specific locations relative to the splicing event, providing a testable mechanistic hypothesis. Our methodology can thus help finding splicing changes that are associated to epigenetic variations and small RNAs between cell types and conditions, thereby providing a mean to uncover new mechanisms of cancer development.

[1] Allo,M. Et al. (2009) Nature structural and molecular biology, 16:7, 717-725.

[2] Luco,R.F. et al (2010). Science, 327, 996-1000.

[3] Ernst,J. and Kellis,M. (2010). Nature Biotechnology, 28, 817-825.

[4] Barash,Y. Et al. (2010). Deciphering the splicing code. Nature, 465, 53-59.

[5] Althammer S, et al. (2011) Bioinformatics. 2011 Oct 12.

Student Symposium Session 1. Talk #2

Stop codons in bacteria are not selectively equivalent.

Inna Povolotskaya[1], Fyodor Kondrashov[1] and Peter Vlasov[1]

1.CRG, Barcelona, Spain

Many global patterns in molecular evolution are defined by the genetical code, including rates of nonsynonymous and synonymous evolution, synonymous codon usage and the optimality of the genetic code. The evolution and usage of stop codons, however, have not been rigorously studied with the exception of coding of non-canonical amino acids. Here, we study the rate of evolution and genomic frequency of TAA, TGA and TAG canonical stop codons in bacterial genomes. We find that stop codons evolve slower than synonymous sites, suggesting the action of weak negative selection. However, the frequency of stop codon usage relative to genomic nucleotide content indicates that this selection regime is not straightforward. The usage of TAA and TGA stop codons is GC-content dependent, with TAA decreasing and TGA increasing with GC content, while TAG frequency is independent of nucleotide content. We thus modeled stop codon usage and nucleotide content with mutation rates and two selection on nucleotide content and TAG frequency as parameters. We found that the relationship between stop codon frequencies and nucleotide content cannot be explained by mutational biases or selection on nucleotide content. However, with weak nucleotide content-dependent selection on TAG, $-0.5 < Nes < 1.5$, the model fits all of the data and recapitulates the lack of a relationship of TAG and nucleotide content. For biologically plausible rates of mutations we show that, in bacteria, TAG stop codon is universally associated with lower fitness, with TAA being the optimal stop codon for G- content < 16% while for G-content > 16% TGA has a higher fitness than TAG.

Student Symposium Session 1. Talk #3

Deep sequencing of RNA from distinct subcellular fractions shows that splicing in the human genome occurs predominantly during transcription.

Hagen Tilgner[1,2], João Curado[1,3] and Roderic Guigó[1]

1 Centre for Genomic Regulation (CRG) and UPF, Barcelona, Spain. 2 Stanford University, Stanford, CA, U.S.A. 3 GABBA, University of Porto, Porto Portugal

Splicing remains an incompletely understood process. Recent findings suggest that chromatin structure participate in its regulation. Here, we analyze the RNA from a number of cellular fractions obtained trough RNASeq in the cell line K562. We show that in the human genome, splicing occurs predominantly during transcription. We introduce the coSI measure, based on the RNASeq reads mapping to exon junctions, to assess the degree of completion of splicing around internal exons. We show that, as expected, for the vast majority of exons splicing is fully completed (coSI close to 1) in cytosolic polyA+ RNA. In contrast, only for about 5% of the exons, splicing is fully completed in the RNA associated to the chromatin (which includes the RNA in the act of being transcribed). However, only a tiny fraction of exons (<0.3%) show highly uncompleted splicing (coSI close to 0) in chromatin associated RNA, indicating that most genes undergo splicing while being transcribed. Consistent with co-transcriptional splicing, we have found significant enrichment of spliceosomal snRNAs in chromatin associated RNA compared to other cellular RNA fractions and other non- spliceosomal snRNAs. CoSI scores show a decreasing 5' to 3' bias, pointing to a "first transcribed, first spliced"-rule, yet more downstream exons carry other characteristics, favoring rapid, co-transcriptional intron removal. Exons with low coSI values—that is, in the process of being spliced—are enriched with chromatin marks, consistent with a role for chromatin in splicing during transcription. Long non-coding-RNAs appear to be spliced later and might remain unspliced in some cases.

Oral presentations

Session 2. **Student Symposium**

Monday 23rd @ 2:15PM

Student Symposium Session 2. Talk #1

Low-complexity regions as a mechanism of protein diversification.

Núria Radó-Trilla[1] and M.Mar Albà[1,2]

1. Evolutionary Genomics Group, Fundació Institut Municipal d'Investigació Mèdica (FIMIM)- Universitat Pompeu Fabra (UPF), Barcelona, Spain. 2. Catalan Institution for Research and Advanced Studies (ICREA), Barcelona, Spain.

Regions of very simple amino acid composition, known as low-complexity regions (LCRs), are surprisingly abundant in protein sequences. Experimental data has shown that the formation of novel LCRs, or the modification of existing ones, can have functional consequences, for example in the localization of the protein [1] or its capability to regulate transcriptional activity [2]. We have compared the LCRs present in a large set of orthologous proteins obtained from Ensembl from five chordate species (human, mouse, chicken, zebrafish and *C. intestinalis*) using the algorithm SEG [3]. Using parameters that ensure that only highly repetitive sequences are recovered, we have obtained that about 10% of mammalian proteins contain at least one LCR, a significantly higher fraction than in chicken or fish (about 7%) or *C. intestinalis* (3.2%). Is the larger number of LCRs in mamamals due increased LCR gain, or to decreased LCR lost? Does the composition of LCRs changes in different vertebrate branches? Which are the functional implications of LCRs?

In order to answer these questions we have followed up the evolutionary history of 1,690 non- redundant LCRs. We have determined the degree of phylogenetic conservation of LCRs using parsimony criteria. We have run SEG with quite relaxed parameters to avoid underestimation of the age of LCRs. With this data we have been able to estimate the number of LCRs gained and lost in each lineage, as well as the number of LCRs in ancestral nodes. We conclude that the number of LCRs has tended to increase in fishes and mammals in relation to their ancestors, but this is not the case of chicken. The most dramatic increase has taken place in the evolution from the common Amniota ancestor to present-day mammals (about 40% increase). Some particular types of LCRs, such as alanine-rich stretches, have increased spectacularly, and, given the role of some of them in transcriptional repression, may have incremented regulatory complexity. Further, we have been able to determine that the main force of the LCR increase in mammals is a lower rate of LCR lost, rather than a higher rate of LCR gain. In the case of fishes the increase in the number of LCRs could be associated with the whole genome duplication (WGD) that took place at the base in this lineage.

To investigate whether LCRs are important players in the acquisition of novel functions by gene duplicates (neofunctionalization), we have measured the LCR content of 154 human transcription factor gene families (transcription factors are specially rich in LCRs) that originated during the two rounds of WGD at the base of the vertebrates. Interestingly, 31% of the duplicated genes in our dataset contain LCRs, but only 17% of the non-duplicated genes do. In addition, LCRs in gene families are often copy-specific and show a significant tendency to have been formed soon after the gene duplication event. This strongly suggests that LCRs are likely to be important in TF gene duplicate neofunctionalization. We are currently performing experiments to confirm this.

[1] Salichs et al. PLoS Genet 5(3):e1000397 (2009)

[2] Fondon et al. PNAS 101 (2004) 18058-18063

[3] Wootton et al. Methods in Enzymology 206 (1996) 554-571

Student Symposium Session 2. Talk #2

SNP4Forensic Project.

Yassine Souilmi[1], Nabil Zaid[1] and Saaid Amazazi[1]

1.Faculty of Sciences, Mohammed V University-Agdal, Rabat, Morocco

We have examined databases (including SNPforID, HapMap and dbSNP) containing SNPs, microsatellites, and mutations. Our interest in SNPs was justified by their growing importance in forensic and also because it allows exploiting DNA damaged fragments.

For this purpose, we have developed an algorithm [SNP4forensic], which when implemented and applied for a particular database [eg, Hap Map], treats billions of information items, filters them by criteria (validity status, GC content, primers test,...), retaining only those that may be useful for forensic usage [SNPs candidates]. This approach considerably facilitates forensic investigation and the development of new identification kits specifics for each ethnic population and therefore more discriminating.

The filtered data were stored in a specific database (DataMart) called SNP4forensic DM. The Datamart contain even the primers of each SNP. This approach makes the research bench easier by reducing considerably the number of candidates to consider, and thus the cost of the study.

Student Symposium Session 2. Talk #3

Homol-MetReS: A web application for integration between molecular systems biology and evolutionary biology.

Hiren Karathia[1], Anabel Usie[2], Ester Vilaprinyo[3], Francesc Solsona[2] , Ivan Teixidó[2], Albert Sorribas[1], and Rui Alves[1]

1. Departament Ciències Mèdiques Bàsiques, Universitat de Lleida & IRBLleida, Lleida, Spain, 2. Department d'Informàtica i Enginyeria Industrial, Universitat de Lleida, Av. Jaume II n°69, 25001 Lleida, Spain. 3. Evaluation and Clinical Epidemiology Department, Hospital del Mar-IMIM, Barcelona, Spain.

In silico reconstruction of biological processes and pathways is facilitated by integrating several levels of functional information that is associated with the proteome of an organism [1]. Thus, a basic need for the reconstruction is the availability of compiled functional classifications of the proteome into various biologically meaningful components. Having such annotation available also facilitates comparing biological circuits between organisms and deciding which model organisms are more adequate to serve as a model to study a given process in a group of other organisms. Recently we have proposed one such method [2]. It is now important to provide an integrated framework based tool that allows any users to apply this method to compare circuits between any numbers of organisms.

In this work we present Homol-MetReS, a user-friendly web application that facilitates: 1) Management of molecular information for each organism, including protein/gene sequences and Functions/processes information assigned to the sequence. 2) (Re)assignment and integration of functional information to proteins or genes as per standard terms defined in GO, EC Numbers, KEGG pathways databases or user defined terms. 3) Creation of organisms- centric clusters of orthologs, homologues and absent genes. 4) Visual comparison between organisms of sets of proteins involved in different specific processes and appropriate choice of model organisms.

We illustrate the functioning of Homol-MetReS with a case study that compares the full proteome of *Saccharomyces cerevisiae* to those of 57 other eukaryotes. This allows us to identifying the organisms for which the set of proteins participating in different processes are more similar to the corresponding set in *S. cerevisiae*.

[1] Alves, R. and A. Sorribas (2007). "In silico pathway reconstruction: Iron-sulfur cluster biogenesis in Saccharomyces cerevisiae." BMC systems biology 1: 10.

[2] Karathia, H., E. Vilaprinyo, et al. (2011). "Saccharomyces cerevisiae as a model organism: a comparative study." PLoS One 6(2): e16015.

Oral presentations

Session 1. **Genome Architecture**
Tuesday 24th @ 10:30AM

Genome Architecture. Talk #1

Joint estimation of methylation probability and genotype from whole genome bisulfite sequence data.

Simon C Heath[1]

1.National Center for Genomic Analysis (CNAG), Barcelona

DNA methylation is an important mechanism in vertebrates for transcription regulation. There are many approaches to measuring methylation with the current ‘gold standard’ being whole-genome bisulfite sequencing, which allows single base resolution of methylation status at (potentially) all cytosines in the genome. Methylation in vertebrates is mainly seen in CpG di-nucleotides, although methylation of isolated cytosines has been reported [1]. The principle of bisulfite sequencing is that bisulfite treatment of DNA converts non-methylated cytosines to uracils, creating an apparent C->T mutation after sequencing of the converted DNA. If the non-converted genome sequence is known, then it is frequency of the C->T changes can be used to estimate the methylation probability for each C. In practice the analysis is complicated by two factors: (a) the difficulty of mapping possibly partially converted reads with the information loss resulting from the C->T changes and (b) sequence variation, so that the unconverted genome is not known a priori.

A method is presented that allows for joint estimation of genotypes (on the unconverted genome) and methylation probabilities for C nucleotides on either DNA strand. The approach uses standard short-read mappers combined with a custom variant caller that performs the joint analysis. The method can combine unconverted and converted sequence reads in the same analysis if both are available to improve the precision of the estimates. In addition to the genotype and methylation estimates, the method framework also provides a means for testing the methylation (and genotype) estimates in single and paired samples. The precision of the methylation estimates are given as output from the analysis, and for paired samples a statistical test for differentially methylated residues is generated. To test the accuracy of the model predictions, a comparison was made between methylation estimates obtained from the Illumina 450k methylome array on a subset (450,000 or 2%) of CpG sites (5 samples), and estimates obtained from whole-genome bisulfite sequencing with the analysis method described here. The two sets of estimates showed very good agreement (r^2>0.97), demonstrating the reliability of the analysis approach.

[1] R. Lister et al. Nature 462 315-322

Genome Architecture. Talk #2

Gene-specific count data distributions are required in RNA-seq experiments with extensive replication.

Mikel Esnaola[1], Pedro Puig[2], David Gonzalez[3], Robert Castelo[4] and Juan Ramón Gonzalez[1,2,5]

1. Center for Research in Environmental Epidemiology (CREAL). 2. Dept. of Statistics, Universitat Autònoma de Barcelona (UAB). 3. Center for Genomic Regulation (CRG). 4. Dept. of Experimental and Health Sciences, Universitat Pompeu Fabra (UPF). 5. Hospital del Mar Research Institute (IMIM).

High-throughput RNA sequencing (RNA-seq) offers unprecedented power to capture the real dynamics of gene expression. Experimental designs with extensive biological replication give a unique opportunity to exploit this feature and distinguish expression profiles with higher resolution. RNA-seq data analysis methods so far have been mostly applied to data sets with few replicates and are based on two well-known count data distributions, the Poisson and the negative binomial.

Here we show, however, that the rich diversity of expression profiles produced by extensively-replicated RNA-seq experiments require additional count data distributions to capture the gene expression dynamics revealed by this technology. We provide a new method for differential expression analysis implemented in a package for R called `tweeDEseq` and based on a broader class of count-data models, that enable making different distributional assumptions on different genes and groups of samples.

We demonstrate that this results in shorter and more accurate lists of differentially expressed genes by surveying the tiny fraction of sex-specific gene expression changes in human lymphoblastoid cell lines. We conclude that the more adequate fit to the underlying biological variability provided by the proposed method will become critical when analyzing RNA-seq samples of larger heterogeneity such as those derived from complex diseases like cancer.

A Role for tRNA Modifications in Genome Structure and Codon Usage.

Eva Maria Novoa[1], Mariana Pavon-Eternod[2], Tao Pan[2] and Lluis Ribas de Pouplana[1,3]

1. Institute for Research in Biomedicine (IRB), c/ Baldiri Reixac 15-21 08028 Barcelona, Spain. 2. Department of Biochemistry and Molecular Biology, University of Chicago, Chicago, IL 60637, USA. 3. Catalan Institution for Research and Advanced Studies (ICREA), Passeig Lluis Companys 23, 08010 Barcelona, Spain.

Despite the central role of tRNAs in protein translation, the connections between tRNA gene population dynamics and genome evolution have rarely been explored. It is known that in unicellular organisms the most abundant codons are recognized by the most abundant tRNAs in the cell [1]. However, we do not understand the reasons for the variability between tRNA pools of different species, nor the principles that determine tRNA gene abundances or genomic codon composition.

Here we report that two specific tRNA wobble base modifications contributed to genome evolution and extant codon usage biases. We have analyzed the genomes of over 500 species in terms of their tRNA gene populations and codon usage. Through this analysis we identify two kingdom-specific modifications whose frequency directly correlates with genome- wide gene expression levels. We show that, contrary to prior observations, genomic codon usage and tRNA gene frequencies correlate in Bacteria and Eukarya if these two modifications are taken into account, and that presence or absence of these modifications explains patterns of gene expression observed in previous studies. Finally we experimentally demonstrate that human gene expression levels correlate well with genomic codon composition if these identified modifications are considered [2].

The discovery of kingdom-specific strategies to optimize translation efficiency opens new possibilities to further improve heterologous gene expression systems. Indeed, heterologous protein expression may be further improved if gene compositions are designed to match the mature tRNA gene population of the host species.

[1] Tuller et al. Cell 141 (2010) 344-354.

[2] Novoa et al. Cell (2011). Accepted.

Genome Architecture. Talk #4

The first HPC pipeline for Next Generation Sequencing data analysis.

Ignacio Medina[1,3], Joaquin Tárraga[1,3] and Joaquín Dopazo[1,2,3]

1.Department of Bioinformatics and Genomics, Centro de Investigación Príncipe Felipe (CIPF), Valencia, Spain. 2. CIBER de Enfermedades Raras (CIBERER), Valencia, Spain. 3. Functional Genomics Node (INB), CIPF, Valencia, Spain

Last years next generation sequencing has proven to be a revolutionary technology. Its amazing throughput allow researchers to sequence and interrogate genomes in a few days for a relatively low price. Since it appeared the sequencing price has not stopped falling while coverage and quality has increased. This technology is changing the way how researchers are conducting their experiments by adapting classical experiments in a NGS version, so we can find exome re-sequencing and variant call analysis to genotype or find new SNPs or variants associated with disease; rna-seq experiments to perform a differential expression analysis or transcript isoforms detection; chip-seq to look for genomic targets for proteins of interest between others.

NGS technology is allowing researchers to obtain very relevant information faster then ever and make new discoveries, but is also challenging bioinformatics because the volume of data has experimented an increase of 1000x going from some hundreds of MegaBytes in a microarray experiment to some hundreds of GigaBytes in NGS. This is making classical data analysis software a limitation in this new scenario in which we can have many TeraBytes of raw sequences and spend several days for analyzing the data. Computer technologies are also evolving very fast, last years we have seen how GPUs processors has increased their power to an unprecedented manner, also new APIs like CUDA for Nvidia GPUs have appeared making easier then ever to develop software that take profit of all this computation power.

Here we present the first HPC pipeline using GPUs processors to analyze NGS raw data. We have focused in every step of exome re-sequencing data analysis pipeline and so far we have implemented in Nvidia CUDA many algorithms and tools covering: a) FASTQ QC and preprocessing, b) Burrows-Wheeler Transform and Smith-Waterman for read mapping, c) SAM/BAM QC and utilities, d) variant calling analysis. Now we are also implementing algorithms and tools to analyze rna-seq or meth-seq data. First results show that we can speed-up the data analysis time by 100x when compared to classical CPU software, going from several hours to minutes of processing. So, we will provide researchers with a complete and integrated set of tools that will allow them to perform analysis in a scale of minutes instead of days.

Structural variation from next generation sequencing. Limits of the technology and lessons from the Great Ape Genome Project.

Tomas Marques-Bonet[1,2]

1.Institució Catalana de Recerca i Estudis Avançats (ICREA). 2. IBE, Institut de Biologia Evolutiva (UPF-CSIC), CEXS, Barcelona, 08003, Spain

Next generation sequencing (or High-throughput sequencing (HTS)) technologies are crucial in the field of genomics, providing unprecedented amount of information that is supposed to resolve the study of genomic variation. Although advances have been made there are still several problems in the understanding of all range of genomic variation from HTS datasets. However, the full spectrum of genetic diversity in human and non-human species is still not well understood given that middle range genomic variation (encompassing repeats and structural variation) is usually hardly identified. To discuss the problems and limitation of this technology, I will use The Great Ape genome project, in which we set out to understand the most complete catalog of genomic variation via next-generation sequencing of a diverse sample of high coverage full genome sequencing of ~100 great ape genomes. The set includes wild-born/unrelated specimens from all major subspecies including 30 chimpanzees, 15 bonobos, 25 gorilla and 16 orangutans. Our analyses have shown that different spectrum of genomic variation contributes more to genetic diversity at the basepair level than single nucleotide variation. Of particular interest, we can now date all gene duplications and determine whether they are fixed or polymorphic between subspecies. The genomic resource we have developed provides a powerful tool to study changes in the frequency and pattern of all forms of genetic during the last 15 million years of human and great ape evolution.

Oral presentations

Session 2. **Highlights 2011 (I)**
Tuesday 24th @ 2:30PM

Highlights 2011 (I). Talk #1

The three-dimensional folding of the α-globin gene domain reveals formation of chromatin globules.

Davide Baù[1], Amartya Sanyal[2], Bryan R Lajoie[2], Emidio Capriotti[1], Meg Byron[3], Jeanne B Lawrence[3], Job Dekker[2] & Marc A Marti-Renom[1]

1 Structural Genomics Unit, Bioinformatics and Genomics Department, Centro de Investigación Príncipe Felipe, Valencia, Spain. 2 Program in Gene Function and Expression, Department of Biochemistry and Molecular Pharmacology, University of Massachusetts Medical School, Worcester, Massachusetts, USA. 3 Department of Cell Biology, University of Massachusetts Medical School, Worcester, Massachusetts, USA.

We developed a general approach that combines chromosome conformation capture carbon copy (5C) with the Integrated Modeling Platform (IMP) to generate high-resolution three- dimensional models of chromatin at the megabase scale. We applied this approach to the ENm008 domain on human chromosome 16, containing the α-globin locus, which is expressed in K562 cells and silenced in lymphoblastoid cells (GM12878). The models accurately reproduce the known looping interactions between the α-globin genes and their distal regulatory elements. Further, we find using our approach that the domain folds into a single globular conformation in GM12878 cells, whereas two globules are formed in K562 cells. The central cores of these globules are enriched for transcribed genes, whereas nontranscribed chromatin is more peripheral. We propose that globule formation represents a higher-order folding state related to clustering of transcribed genes around shared transcription machineries, as previously observed by microscopy.

Highlight of article Nature Structural & Molecular Biology 18, 107–114 (2011).

Highlights 2011 (I). Talk #2

Chromatin organization in sperm may be the major functional consequence of base composition variation in the human genome.

Tanya Vavouri[1] and Ben Lehner[1]

1 EMBL-CRG Systems Biology Unit, Centre for Genomic Regulation, Universitat Pompeu Fabra, Barcelona, Spain

Chromatin in sperm is different from that in other cells, with most of the genome packaged by protamines not nucleosomes. Nucleosomes are, however, retained at some genomic sites, where they have the potential to transmit paternal epigenetic information. It is not understood how this retention is specified. Here we show that base composition is the major determinant of nucleosome retention in human sperm, predicting retention very well in both genic and non-genic regions of the genome. The retention of nucleosomes at GC-rich sequences with high intrinsic nucleosome affinity accounts for the previously reported retention at transcription start sites and at genes that regulate development. It also means that nucleosomes are retained at the start sites of most housekeeping genes. We also report a striking link between the retention of nucleosomes in sperm and the establishment of DNA methylation-free regions in the early embryo. Taken together, this suggests that paternal nucleosome transmission may facilitate robust gene regulation in the early embryo. We propose that chromatin organization in the male germline, rather than in somatic cells, is the major functional consequence of fine-scale base composition variation in the human genome. The selective pressure driving base composition evolution in mammals could, therefore, be the need to transmit paternal epigenetic information to the zygote.

Highlight of article PLoS Genet 7(4): e1002036.

Highlights 2011 (I). Talk #3

Distributed biological computation with multicellular engineered networks.

Sergi Regot[1], Javier Macia[2], Núria Conde[1,2], Kentaro Furukawa[3], Jimmy Kjellén[3], Tom Peeters[1], Stefan Hohmann[3], Eulàlia de Nadal[1], Francesc Posas[1], Ricard Solé[2,4,5].

1 Cell signaling unit, Departament de Ciències Experimentals i de la Salut, Universitat Pompeu Fabra (UPF), E-08003 Barcelona, Spain 2 ICREA-Complex Systems Laboratory, Universitat Pompeu Fabra (UPF), E-08003 Barcelona, Spain 3 Department of Cell and Molecular Biology/Microbiology, University of Gothenburg, Box 462, 40530 Gothenburg, Sweden 4 Santa Fe Institute, Santa Fe, New Mexico 87501, USA 5 Institut de Biologia Evolutiva, CSIC-UPF, Passeig Maritim de la Barceloneta, 37-49, 08003 Barcelona, Spain

Ongoing efforts within synthetic and systems biology have been directed towards the building of artificial computational devices using engineered biological units as basic building blocks. Such efforts, inspired in the standard design of electronic circuits, are limited by the difficulties arising from wiring the basic computational units (logic gates) through the appropriate connections, each one to be implemented by a different molecule. Here, we show that there is a logically different form of implementing complex Boolean logic computations that reduces wiring constraints thanks to a redundant distribution of the desired output among engineered cells. A practical implementation is presented using a library of engineered yeast cells, which can be combined in multiple ways. Each construct defines a logic function and combining cells and their connections allow building more complex synthetic devices. As a proof of principle, we have implemented many logic functions by using just a few engineered cells. Of note, small modifications and combination of those cells allowed for implementing more complex circuits such as a multiplexer or a 1-bit adder with carry, showing the great potential for re-utilization of small parts of the circuit. Our results support the approach of using cellular consortia as an efficient way of engineering complex tasks not easily solvable using single-cell implementations.

Highlight of article Nature. 2011 Jan 13;469(7329):171-2.

Oral presentations

Session 3. **Structural Bioinformatics**
Tuesday 24th @ 4:30PM

Molecular recognition of SH2-phosphopeptide by molecular dynamics

Toni Giorgino[1], Ignasi Buch[1], and Gianni De Fabritiis[1]

1. Computational Biochemistry and Biophysics Laboratory (GRIB-IMIM), Universitat Pompeu Fabra, Barcelona, Spain

Src homology 2 (SH2) domains are small (~100 amino acids) and well-conserved protein domains which recognize, with high affinity and specificity, short amino acid sequences (3 to 6 residues) introduced by a phosphorylated tyrosine, pY(+0). Approximately 110 proteins in the human genome contain a SH2 domain, highlighting their prominent role in signal transduction.

In this work we computed a set of trajectories via high-throughput unbiased MD simulations in explicit solvent on initially unbound pYEEI to elucidate the pathway leading to association with SH2 [1]. An ensemble of 772 trajectories of 200 ns were started with the ligand placed at least 30 Å away from the binding site. After 200 ns, five out of the 772 trajectories evolved in a bound configuration, defined as ligand's backbone RMSD < 2 Å with respect to the 1LKK crystal. An order-of-magnitude estimate can be obtained assuming first-order kinetics, which yields $k_a \sim 16 \times 10^5$ M^{-1} s^{-1}; the value is within the range of those obtained in vitro on various ligands.

We report the common molecular pathways of the five spontaneous binding events between the p56 lck SH2 and the pYEEI peptide, observed at atomistic detail. First, the peptide undergoes diffusion guided by electrostatic forces, which yield a transient ensemble of contacts within tens of nanoseconds. The established contacts restrict the ligand search space, until the two major events implied by a characteristic "two-pronged" SH2 recognition model [2] occur: first, the phosphorylated moiety is buried into the native binding pocket; second, the C-terminus of the peptide falls into the hydrophobic pocket. The hydrophobic terminus can only snap in its native position when a gate formed by residues Ser (EF1) and Leu (BG4) is in its open state [3]. The native conformation is finally achieved when the BC loop closes over the phosphorylated tyrosine, thus stabilizing the complex.

In conclusion, we have shown how, in a system with moderate structural complexity and a flexible ligand, the molecular determinants and their temporal dependencies could be revealed. Further detailed quantification of the kinetic determinants of binding, allowed by methodologies based on ensembles of trajectories [4], may enable drug design processes encompassing the rational optimization of kinetic parameters.

[1] Buch, I.; Harvey, M. J.; Giorgino, T.; Anderson, D. P.; De Fabritiis, G. Journal of Chemical Information and Modeling 2010, 50, 397-403.

[2] Waksman, G.; Shoelson, S. E.; Pant, N.; Cowburn, D.; Kuriyan, J. Cell 1993, 72, 779-790.

[3] Taylor, J. D.; Ababou, A.; Fawaz, R. R.; Hobbs, C. J.; Williams, M. A.; Ladbury, J. E. Proteins 2008, 73, 929-940.

[4] Buch, I.; Giorgino, T.; De Fabritiis, G. Proc. Natl. Acad. Sci. U.S.A. 2011.

A global network of crosstalking post-translational modifications

Pablo Minguez[1], Luca Parca[2], Francesca Diella[1,3], Daniel Mende[1], Runjun Kumar[4], Manuela Helmer- Citterich[2], Anne-Claude Gavin[1], Vera van Noort[1] and Peer Bork[1,5]

1. European Molecular Biology Laboratory, Meyerhofstrasse 1, 69117 Heidelberg, Germany. 2. Department of Biology, Centre for Molecular Bioinformatics, University of Rome 'Tor Vergata', Via della Ricerca Scientifica snc, 00133 Rome, Italy. 3. Molecular Health GmbH, Belfortstrasse, 2, 69115 Heidelberg, Germany. 4. Washington University, St. Louis, Missouri, United States of America.. 5. Max-Delbruck-Centre for Molecular Medicine, Berlin-Buch, Germany

Protein function is partially regulated by the tight interplay between protein-protein interactions and protein post-translational modifications (PTMs) [1]. These modifications cover a significant proportion of eukaryotic proteins [2,3], exist in a large variety of forms (PTM types) and can be present in different combinations of several types [4]. The examples of histone tail modifications [2] and p53 regulation [3] have been used to suggest that most eukaryotic proteins may be subjected to PTM co-regulation and a general PTM code has been postulated [5]. Aiming at a more global view of the interplay between PTM types, we collected non-redundant residue positions for more than 115,000 experimentally verified modification sites of 13 frequent PTM types in 8 eukaryotes, compared the speed of evolution of the types to each other (carboxylation being the most conserved and SUMOylation the least) and developed a method for measuring PTM crosstalk within proteins based on mutual information [6] of the co-occurrence of sites across eukaryotes. We found that all PTM types crosstalk with a large number of other types forming a global network, evaluated under a statistical framework, that comprise in human alone more than 50,000 residues in about 6,000 proteins. Several pairs of PTM types tend to crosstalk more if they are close in sequence or structural proximity implying a direct physical interaction between the modified sites. In addition to well-studied PTM type interplay in nucleus and cytosol, we found substantial PTM type crosstalk within proteins that are secreted or membrane-associated with a high proportion of links no previously described in the literature. Analysis of PTM type crosstalk also revealed novel functional associations with protein domains and protein short linear motifs which can regulate protein binding [7]. Although some of these kind of associations (protein domains) have been reported before for single modifications [1], the present work represents the first systemic and proteome-wide analysis of co-regulating modifications linked to protein function. The global network of crosstalking PTM types implies a complex and intertwined post-translational regulation landscape that is likely to regulate multiple functional states and location of many if not all eukaryotic proteins.

[1] B.T. Seet et al. Nature reviews. Molecular cell biology 7, 473-83 (2006). [2] P. Cohen. Trends in biochemical sciences 25, 596-601 (2000).

[3] B.T. Weinert et al. Science Signaling 4, ra48-ra48 (2011).

[4] X.-J. Yang. Oncogene 24, 1653-62 (2005).

[5] B. A. Benayoun, R. A. Veitia, Trends in cell biology. 19, 189-97 (2009).

[6] T. M. Cover, J. A. Thomas. John Wiley & Sons, New York (1991).

[7] V. Neduva et al. PLoS biology 3, e405 (2005).

Structural Bioinformatics. Talk #3

To Bind or not To Bind: Predicting protein-protein interactions from favouring and disfavouring local structural features.

Joan Planas-Iglesias[1], Jaume Bonet[1], Manuel Marín-López[1], Elisenda Feliu[1], and Baldo Oliva[1].

1. Structural Bioinformatics Lab. GRIB. Universitat Pompeu Fabra.

INTRODUCTION: Protein-protein interactions (PPIs) are crucial to understand how proteins perform their cellular functions. Therefore, correctly identifying the PPI network (or protein interactome) of a given organism is useful not only to shed light on the key molecular mechanisms behind a biological function but also to infer the functionality of a protein based on its interactions.During the past ten years, Margalit and coworkers have performed a series of non-docking experiments to predict PPIs [1-3]. Using the knowledge of experimentally determined PPIs they initially developed an association method to describe pairs of correlated domains found in PPIs.

OBJECTIVE: It is striking the differences between the precision of these results and the ones achieved by docking experiments. It has been proposed that docking poses may be reflecting loose contacts that would allow the correct reorientation of the interacting partners. This interaction model suggests that several regions of the protein, may be related to the molecular association, a central concept of the funnel-like intermolecular energy landscape used to describe PPIs [4]. To elucidate whether determinants of protein interaction are local to the interacting region or not, in this work we use an association method that takes advantage of both the negatome and known PPIs to define characteristic features for known interacting and non-interacting protein pairs.

MATERIAL AND METHODS: Using BIANA [5] we extract all available Y2H PPIs as positive model of interaction. Conversely, negative data is obtained from Negatome [6]. We apply an association method to these data to extract characteristic signatures of both sets. These are used to score potential protein interacting pairs. We evaluated the performance of positive and negative signatures, and their combination (log2 ratio), as predictors of PPIs. We also analyzed the role of the number of positive and negative signatures, in the context of the funnel-like intermolecular energy landscape, as predictors of PPIs.

RESULTS: Our results show that, while positive signatures can be used to predict interacting domains, it is negative signatures, which have an important role on using loops for PPI prediction. Furthermore, in all cases it is the combination of both positive and negative signatures the best predictor.

CONCLUSIONS: Our results strongly suggest that it is the balance between typical interacting and non-interacting structural features in the protein surface which determine if a pair of proteins will interact or not.

[1] Akiva, E. et al. Proc Natl Acad Sci U S A 105 (2008) 13292-7.

[2] Sprinzak, E. et al. Proc Natl Acad Sci U S A. 103 (2006) 14718-23.

[3] Sprinzak, E. et al. Journal of molecular biology 311(2001) 681-92.

[4] Wass, M.N. et al. Molecular systems biology 7 (2011) 469.

[5] Garcia-Garcia, J. et al. BMC bioinformatics 11 (2010) 56.

[6] Smialowski, P. et al. Nucleic Acids Res. 38 (2010) D540-4.

Structural Bioinformatics. Talk #4

A reverse combination of structure-based and ligand-based strategies for virtual screening.

Álvaro Cortés-Cabrera[1,2], Federico Gago[1], and Antonio Morreale[2]

1. Departamento de Farmacología, Universidad de Alcalá, E-28871 Alcalá de Henares, Madrid, Spain. 2. Unidad de Bioinformática, Centro de Biología Molecular Severo Ochoa (CSIC/UAM), Campus UAM, c/ Nicolás Cabrera 1, E-28049 Madrid, Spain.

A new approach is presented that combines structure- and ligand-based virtual screening in a reverse way. Opposite to the majority of the methods, a docking protocol is first employed to prioritize small ligands ("fragments") that are subsequently used as queries to search for similar larger ligands in a database. For a given chemical library, a three-step strategy is followed consisting of (i) contraction into a representative, nonredundant set of fragments; (ii) selection of the three best-scoring fragments docking into a given macromolecular target site; and (iii) expansion of the fragments' structures back into ligands by using them as queries to search the library by means of fingerprint descriptions and similarity criteria. We tested the performance of this approach on a collection of fragments and ligands found in the ZINC database [1] and the Directory of Useful Decoys [2], and compared the results with those obtained using a standard docking protocol. The new method, which has been implemented in our virtual screening data management on an integrated platform (VSDMIP [3,4]), provided better overall results and was several times faster. We also studied the chemical diversity covered by both methods, and concluded that the novel approach covered it almost entirely but at a much smaller computational cost.

[1] JJ. Irwin J. Chem. Inf. Model. 45 (2005) 177.

[2] N. Huang et al. J. Med. Chem. 49 (2006) 6789.

[3] R. Gil-Redondo et al. J. Comput.-Aided Mol. Des. 23 (2009) 171.

[4] A. Cortés-Cabrera et al. J. Comput.-Aided Mol. Des. 25 (2011) 813.

Oral presentations

Session 4. **Genome annotation**

Wednesday 25th @ 10:30AM

Genome Annotation. Talk #1

Sma3s: a 3 stages software for sequences make sense.

Antonio M. Mérida[1], Gonzalo M. Claros[2], Oswaldo Trelles[1] and Antonio J. Perez[3]

1 Computer Architecture Department, University of Malaga, Spain. 2 Biochemistry department, University of Malaga, Spain. 3 (CABD), CSIC-UPO. Experimental Science Faculty (Genetics area), University Pablo de Olavide, Spain

Abstract: We present Sma3s, a biological sequence annotation tool especially focused in the massive annotation of sequences obtained either from any kind of gene library or genome. Sma3s tool is composed of 3 modules that sequentially solve the annotation from: (a) already existing well-annotated sequences (virtually the same sequence that the query one); (b) orthologous sequences and (c) groups of sequences sharing statistical significant patterns (the most original module). Sma3s is also able to add annotations to unannotated sequences in the database which make evident the increased accuracy with respect to other similar softwares. As biological descriptors associated to each sequence, Sma3s provides (i) gene ontology terms, (ii) Swiss-Prot *keywords* and (iii) *pathways*, (iv) and InterPro domains, though it can use new annotation types in an easy way.

Introduction: Biological sequence annotation is the process of finding, recovering and incorporating relevant biological information available in public databases regarding to an individual or massive collection of sequences. Traditionally, methods to recover information from sequences take advantage of homology-based searches and infer the annotations simply from the best hit or hits, —*e.g.* AutoFact [1], Blast2GO [2] and GOtcha [3] — or offer a list of the bests scored sequence alignments without specific annotations, *e.g.* ESTAnnotator [4] and EST-PAC [5]. Sma3s was trained with several random sequence sets and provides high levels of prediction accuracy with minimal human participation and computational resources, Sma3s has been implemented in Perl language and uses blast and blastclust for sequence clustering procedures. As main source of data Sma3s uses the UniProt database in plain-text format (*.dat) and specifically the taxonomic division to which the organism under study belongs, which can be easily select them from a FTP server.

Results: The Sma3s settings were tested with the EST library of tomato (*Solanum lycopersicum*) used in the microarray TOM2 which contains 11,461 sequences. Using Sma3s, 8989 were annotated, which corresponds to a 78% of the original dataset and in average with other annotation sets the program shows a sensitivity of 65% and a specificity of 70%. In addition, each putative annotation is together with a probability value which helps to assign the annotation with accuracy.

Acknowledgements: This work has been partially funded by INB-ISCIII and AT2009-0025.

[1] Koski LB. et al. BMC Bioinformatics (2005) Jun 16:6(1):151

[2] Conesa A. et al. Int J Plant Genomics. (2008) 2008:619832

[3] Martin DM. et al. BMC Bioinformatics. 2004 Nov 18;5:178.

[4] Hotz-Wagenblatt A. et al. Nucleic Acids Res. 2003 Jul 1;31(13):3716-9.

[5] Strahm Y. et al. Source Code Biol Med. 2006 Oct 12;1:2.

Genome Annotation. Talk #2

The GEM toolkit: world-class short read mapping, 100% made-in-Spain.

Santiago Marco Sola[1] and Paolo Ribeca[1]

1 Centro Nacional de Análisis Genómico (CNAG), Barcelona, Spain

The GEM project was born in 2008 to provide alignment capabilities to the first Spanish Illumina Genome Analyzer located at the Centro de Regulació Genòmica (CRG) in Barcelona. After several years of continued careful design and optimization of novel algorithms, the GEM tools for short-read analysis have now garnered a significant user base in many research centers around the world. They also represent the workhorse of short-read processing at the Centro Nacional de Análisis Genómico (CNAG), the Spanish national sequencing center located in Barcelona, which is equipped with a peak sequencing capacity of more than 600 Gbases/day.

Here we present the upcoming stable release of the GEM mapper. Not only the GEM mapper is one of the most precise and versatile aligners available so far (it allows for exhaustive variable-depth searches that do no miss any existing match), making it the ideal tool for high-precision genomic studies; it also delivers breathtaking performance, being several times faster than other reference programs in the field (like BWA [1] or Bowtie [2]). We argue that both such capabilities are actually necessary to be able to cope successfully with the bewildering forthcoming increase in short-read sequencing capacity which is expected to be happening soon.

[1] Fast and accurate short read alignment with Burrows-Wheeler Transform. Bioinformatics 25 (2009).

[2] Ultrafast and memory-efficient alignment of short DNA sequences to the human genome. Genome Biology 10 (2009).

Genome Annotation. Talk #3

Highly efficient pre-processing of NGS reads and identification of full-length genes.

Darío Guerrero-Fernández[1], Noé Fernández-Pozo[2], Almudena Bocinos[1], Rocío Bautista[1] and M. Gonzalo Claros[1,2]

1 Plataforma Andaluza de Bioinformática-Centro de Supercomputación y Bioinformática, Universidad de Málaga, Málaga. 2 Departamento de Biología Molecular y Bioquímica, Facultad de Ciencias, Universidad de Málaga, Málaga.

The advent of the technologies so-called next-generation sequencing (NGS) is giving the ability to obtain large amounts of sequences. However, NGS reads are not clean and it is necessary to remove (depending on the experimental approach) polyA/T, adaptors, contaminations, low quality portions, low complexity segments, artefactual duplicates, tags or sample identifiers, etc. Therefore, there is a need for new, fast, efficient, reliable, easy- to-install, user-friendly, pre-processing software for NGS reads for a wide range of computers and experimental conditions (e.g. de novo assembling, mapping, amplicons). SeqTrimNext has been developed to fill these necessities and to cope with all NGS peculiarities, including parallelisation, managing of paired-end reads, managing and grouping sequences by barcodes or tags, and providing output files that can be used as input for downstream analyses. The modular architecture of SeqTrimNext, based on a pipeline of orthogonal plugins, is especially suitable for addition, removal, or reordering of plugins and easy adaptation for future evolution. It also provides a detailed statistics on the input and output datasets and a PDF report where users can find clues for the their sequence qualities. SeqTrimNext recovers more true paired-ends than others (e.g. Newbler), is released with a customised database for contaminants (even if users can use their own contamination database), and provides pre-coded configuration files for the most common NGS analyses. Assembly of SeqTrimNext-treated reads takes less time, provide less but longer contigs (net increase of N50), dramatically diminished numbers of repeated targets and multiply mapped reads, reduce de amount of chimerical contigs and reduces the subsequent time of manual curation of assemblies and mappings because the obtained results lack misconnections due to artefactual sequences. It can be used at http://www.scbi.uma.es/seqtrimnext

When NGS reads were obtained from transcriptomics experiments, the assembly would require a deeper analysis to know its accuracy and reliability, as well as if any full-length gene has been reconstructed and the presence of putative new genes. This can be achieved by means of FullLengtherNext. Using customised DNA and protein databases (FullLengtherNext is provided with a script for constructing the user-adapted database for the taxonomic group of interest), it is able to predict the completeness of sequences (full- length, internal, N-terminal or C-terminal), predict the artefactual contigs (e.g., both strands seems to code for the same or different genes), corrects putative indels to obtain the most probable ORF, and annotates the sequence with the description of the most similar subject with a reliable description (avoiding when possible the «unknown» or «predicted» tags). It works in parallelised or distributed systems and can be tested at http://www.scbi.uma.es/fulllengther2.

Genome Annotation. Talk #4

Relational Learning-based Extension for Reactome Pathways with Sequence Features and Interactions.

Beatriz García-Jiménez[1], Tirso Pons[2], Araceli Sanchis[1], and Alfonso Valencia[2]

1 Computer Science Department, Universidad Carlos III de Madrid, Madrid, Spain. 2 Structural Biology and BioComputing Programme, Spanish National Cancer Research Centre (CNIO), Madrid, Spain.

Biological pathways are an important component of systems biology. During the past decade, an increasing number of pathway databases have been established to document the expanding knowledge regarding complex cellular processes [1]. As more genome-sequence data become available and a large fraction of these are functionally uncharacterized, it is of interest to explore the possibilities of expanding pathways with potentially related proteins, but not originally included. We have developed a Relational Learning-based Extension (RLE) system for extending cellular pathways. Our approach is related to a set of methods designed for function prediction by means of combinations of simple properties associated to each protein, used in different scenarios [2]. RLE searches for specific proteins molecularly similar to different pathway fragments, instead of proteins with common characteristics in overall pathway at process level.

The predictions are based mainly on sequence features, including the number of isoforms, but also based on some data related with protein-protein interaction networks and protein complexes, i.e. interaction partners with their corresponding sequence features. This relational information makes this system different from others based only on individual characteristics. RLE extends pathways based on a relational representation [3], which lets it apply a combination of relational and propositional machine learning algorithms. Using RLE, we extend 28 human pre-defined Reactome pathways with 383 proteins, not previously annotated in Reactome, and non-sequence redundant with the annotated ones. As expected, our extension differs from that achieved by Glaab et al. [4], a distinct method based only on interaction networks. RLE provides more diversity of functions, not searching only in the proximal space (it means, for example, less predicted proteins in the intersection of pathways).

New putative pathway components predicted by RLE provide useful explanatory information for some of these extended cellular processes, particularly interesting in Electron transport chain, Telomere maintenance and Integrin cell surface interactions. The UniProt annotations and literature findings, when combined with RLE results, increase reliability of the assignment of a protein to a pathway. For example, three of the predicted proteins in Telomere maintenance pathway have been related to ubiquitin conjugation by UniProt annotations. Besides, the three proteins have a RING-finger domain, which is essential for the physical interaction between ubiquitin protein and human telomerase, during the telomere degradation. This process is abolished by a mutation in the RING-finger domain [5]. Therefore, these domain annotations and literature findings do add biological evidences to support our de novo predictions.

[1] Cerami et al. Nucl. Acids Res., 39(suppl 1) (2011), D685–D690.

[2] Jensen et al. Bioinformatics, 19(5) (2003), 635–642.

[3] Dzeroski and Lavrac. Relational Data Mining (2001).

[4] Glaab et al. BMC Bioinformatics, 11(1) (2010), 597.

[5] Kim et al. Genes and development, 19(7) (2005), 776–781.

Oral presentations

Session 5. **Highlights 2011 (II)**

Wednesday 25th @ 2:30PM

Interactome mapping suggests new mechanistic details underlying Alzheimer's disease.

Montserrat Soler-López[1], Andreas Zanzoni[1], Ricart Lluís[1], Ulrich Stelzl[2] and Patrick Aloy[1,3]

1 Institute for Research in Biomedicine, Joint IRB-BSC Program in Computational Biology, 08028 Barcelona, Spain. 2 Max-Planck Institute for Molecular Genetics, 14195 Berlin, Germany. 3 Institució Catalana de Recerca i Estudis Avançats (ICREA), 08010 Barcelona, Spain

Recent advances toward the characterization of Alzheimer's disease (AD) have permitted the identification of a dozen of genetic risk factors, although many more remain undiscovered. In parallel, works in the field of network biology have shown a strong link between protein connectivity and disease. In this manuscript, we demonstrate that AD-related genes are indeed highly interconnected and, based on this observation, we set up an interaction discovery strategy to unveil novel AD causative and susceptibility genes. In total, we report 200 high-confidence protein–protein interactions between eight confirmed AD-related genes and 66 candidates. Of these, 31 are located in chromosomal regions containing susceptibility loci related to the etiology of late-onset AD, and 17 show dysregulated expression patterns in AD patients, which makes them very good candidates for further functional studies. Interestingly, we also identified four novel direct interactions among well- characterized AD causative/susceptibility genes (i.e., APP, A2M, APOE, PSEN1, and PSEN2), which support the suggested link between plaque formation and inflammatory processes and provide insights into the intracellular regulation of APP cleavage. Finally, we contextualize the discovered relationships, integrating them with all the interaction data reported in the literature, building the most complete interactome associated to AD. This general view facilitates the analyses of global properties of the network, such as its functional modularity, and triggers many hypotheses on the molecular mechanisms implicated in AD. For instance, our analyses suggest a putative role for PDCD4 as a neuronal death regulator and ECSIT as a molecular link between oxidative stress, inflammation, and mitochondrial dysfunction in AD.

Highlight of article Genome Res. 2011 21: 364-376.

Highlights 2011 (II). Talk #2

The RAS Superfamily of signaling proteins: a 2011 update.

Gloria Fuentes[1], Antonio Rausell[2], Alfonso Valencia[2] and Ana M. Rojas[3]

1 Biomolecular Modeling and Design Division. Bioinformatics Institute A*STAR. Singapore. 2 Structural Biology and Biocomputing Programme. Spanish National Cancer Research Centre (CNIO). 3 Computational Cell Biology Group. Institute for Predictive and Personalized Medicine of Cancer (IMPPC).

The Ras Superfamily is a perplexing case of functional diversification in the context of a preserved structural framework and a prototypic GTP binding site. Thanks to the availability of complete genome sequences of species representing important evolutionary branch points, we have analyzed the composition and organization of this Superfamily at a greater level of than was previously possible. Phylogenetic analysis of gene families at the organism and sequence level revealed complex relationships between the evolution of this protein Superfamily sequences and the acquisition of distinct cellular functions. Gathering sequence and structural information together, along with advances in computational methods, has helped to identify features important for the recognition of molecular partners and the functional specialization of different members of the Ras.

The cellular organization and signaling in a cell is heavily influenced by the Ras Superfamily of small GTP-binding proteins. In structural terms, the domain contains five loops (G1-G5) that form the nucleotide-binding site, with an interface that is responsible for nucleotide specificity and affinity, and that regulates GTP hydrolysis. The relationships between these proteins and their effectors have been analyzed using distinct phylogenetic approaches in organisms of interest. Although incorporating structural information to evolutionary frameworks have been useful in other signaling proteins, a comprehensive analyses including many representative species and an orthogonal approach to merge information has not been conducted in the Ras Superfamily.

We have updated the most recent version of the human Ras proteins, and compiled the orthologues in 12 complete (or almost complete) genomes representing crucial evolutionary time points. Analysis of the GTP domain of these proteins was conducted using probabilistic inference. Additionally, we compared and contrasted the families to identify the residues (Specificity Determining Positions, SDPs) in the G-domain and to determine any differences that may underlie their specific interactions with effectors using unsupervised approach that is based on Multiple Correspondence Analysis. Then, we analyzed the distribution of the SDPs residues at the interfaces with the effectors, GEF and GAP proteins to identify potential SPDs involved in the specific function.

Our work of the phylogenetic classification of the Ras Superfamily has provided useful data relating to the classification of a set of more divergent members, that can be tentatively classified as independent families. Moreover, the analysis of SDPs provided independent evidence for this classification, as some members behave as independent groups. We discuss these findings at the level of RAS-effector complexes.

Highlight of article JCB (2011) *in press*.

Highlights 2011 (II). Talk #3

Estimation of alternative splicing variability in human populations.

Mar Gonzàlez-Porta[1,2], Miquel Calvo[3], Michael Sammeth[1], and Roderic Guigó[1,4]

1 Bioinformatics and Genomics, Center for Genomic Regulation (CRG) and UPF, Barcelona, Catalonia, Spain. 2 Current affiliation: European Molecular Biology Laboratory, European Bioinformatics Institute, Hinxton, UK. 3 Departament d'Estad ́ıstica, Facultat de Biologia, Universitat de Barcelona (UB) Barcelona, Catalonia, Spain. 4 Departament de Cì encies Experimentals i de la Salut, Universitat Pompeu Fabra, Barcelona, Catalonia, Spain.

DNA arrays have been widely used to perform transcriptome wide analysis of gene ex- pression and many methods have been developed to measure gene expression variability and to compare gene expression between conditions. As RNA-seq is also becoming increasingly popular for transcriptome characterization, the possibility exists for further quantification of individual alternative transcript isoforms, and therefore for estimating the relative ratios of alternative splice forms within a given gene. Changes in splicing ratios, even without changes in overall gene expression, may have important phenotypic effects. Here we have developed statistical methodology to measure variability in splicing ratios within conditions, to compare it between conditions and to identify genes with condition specific splicing ratios. Furthermore, we have developed methodology to deconvolute the relative contribution of variability in gene expression vs variability in splicing ratios to the overall variability of transcript abundances. As a proof of concept, we have applied this methodology to estimates of transcript abundances obtained from RNA-seq experiments in lymphoblastoid cells from Caucasian and Yoruban individuals. We have found that protein coding genes exhibit low splicing variability within populations, with many genes exhibiting constant ratios across individuals. When comparing these two populations, we have found that up to 10% of the studied protein coding genes exhibit population-specific splicing ratios. We estimate that about 60% of the total variability observed in the abundance of transcript isoforms can be explained by variability in transcription. A large fraction of the remaining variability can likely result from variability in splicing. Finally, we also detected that variability in splicing is uncommon without variability in transcription.

Highlight of article Genome Res. published online November 23, 2011.

Oral presentations

Session 6. **Phylogenetics and evolution**

Wednesday 25th @ 4:30PM

Phylogenetics and Evolution. Talk #1

Detecting bacterial interactions from environmental samples: Ecological aggregations favor bacterial cosmopolitanism.

Alberto Pascual-García[1], Javier Tamames[2] and Ugo Bastolla[1]

1 Centro de Biología Molecular (CSIC-UAM), Madrid, Spain. 2 Centro Nacional de Biotecnología (CSIC-UAM), Madrid, Spain.

Microbes are remarkably cosmopolitan, since they are found in very diverse environmental conditions and communities [1]. This property is striking from an ecological point of view, apparently contrasting with the niche view of biodiversity based on competitive exclusion.

Another ecological peculiarity of microbes is the huge biodiversity of their communities. In this work, we use large scale data [1] on presence-absence of bacterial genera in order to reconstruct their ecological interactions. For this task we adopt a recent null model [2] with parameters estimated through maximum likelihood, generalized to account for non random associations between taxons and environments.

Using this model, we analytically assess the significance of co-occurrence or exclusion of bacterial taxons, adopting an objective criterion to establish comparable significance thresholds for aggregation and segregation. We find that taxon aggregation is more frequent than segregation in the bacterial world, which suggests that mutualistic interactions are more common than competitive ones. The resulting aggregation network is found to be significantly clustered and nested, in analogy with mutualistic networks of plants and pollinators [3], whose nestedness has been suggested to favor biodiversity [4]. Interestingly, we find a positive relationship between the number of aggregations and the cosmopolitanism of bacterial taxons. We also observe that phylogenetic relatedness favors aggregation and disfavors segregation, or it makes impossible to assess it, since closely related species either are only found in different environments or significantly aggregate. This suggests that bacterial speciation may have two outcomes, either the adaptation to different environments as in the classic niche view, or the establishment of strong mutualistic cooperation that overcomes competitive exclusion. Therefore, these results suggest that mutualistic associations are quite common in the microbial world, and they are the key to explain the remarkable cosmopolitanism of bacteria and the diversity of their communities.

[1] Javier Tamames, Juan José Abellán, Miguel Pignatelli, Antonio Camacho, and Andrés Moya. Environmental distribution of prokaryotic taxa. BMC Microbiol, 10:85, (2010).

[2] JA Navarro-Alberto and BFJ Manly. Null model analyses of presence absence matrices need a definition of independence. Population Ecology, 51:4, (2009) 505-512.

[3] Jordi Bascompte, Pedro Jordano, Carlos J Mellán, and Jens M Olesen. The nested assembly of plant-animal mutualistic networks. PNAS, 100:16, (2003) 9383-9387.

[4] Ugo Bastolla, Miguel A Fortuna, Alberto Pascual-García, Antonio Ferrera, Bartolo Luque, and Jordi Bascompte. The architecture of mutualistic networks minimizes competition and increases biodiversity. Nature, 458:7241 (2009) 1018-1020.

Does Nature Play Dice with Genomes?

François Serra[1], Verónica Becher[2], and Hernán Dopazo[1]

1 Evolutionary Genomics Laboratory. Bioinformatics and Genomics Department. Centro de Investigación Príncipe Felipe. Valencia. Spain. 2. Computation Department. Facultad de Ciencias Exactas y Naturales. Universidad de Buenos Aires. Argentina

Is there a common rule governing species abundance and diversity (SAD) in ecology and genomics? Even, in the case that such rule exists, it is the same for functional and non- functional components of genomes? It is the same for small and large genomes? To what extent SAD reflects adaptive or stochastic outcomes? Species diversity and their relative abundance have always intriguing ecologists. Ecological models of SAD are descriptive (statistical-based) or mechanistic (niche-based or neutrals). Most mechanistic models assume niche differences as the main cause driving community composition. Neutral models however, consider niche differences among species irrelevant. The unified neutral theory of biodiversity and biogeography (UNTB) [1] is a neutral- stochastic theory originally inspired in neutral population genetic models. UNTB assumes interactions among tropically similar species equivalent on an individual "per capita" basis. This provocative assumption means that these individuals, regardless of the species, appear to be controlled by similar birth, death, dispersal, and speciation rates. Since each species follows a random walk, biodiversity composition emerges randomly in the community [2]. Here, taking advantage of UNTB formulation and the general framework posed by ecological genomics [3, 4] we ask for the relative SAD of genetic elements in genomes. Ensembl database was used to obtain complete functional and non-functional elements of ~500 chromosomes in 30 eukaryote genomes. UNTB+[5] was used to test neutrality and to fit by maximum likelihood UNTB parameters. Ideal models for ecological genomics would consider all diversity of elements populating genomes: satellites sequences, DNA-transposons, LTR-retrotransposons, LINES, SINES, miRNA, scRNA, rRNA, tRNA, genes, and pseudogenes among many other functional and non-functional elements. After ML adjustment of UNTB parameters we found that most chromosomes follow relative SAD according to the expected by UNTB. Functional and non- functional genome elements showed equivalent SAD describing neutral dynamics in chromosomes. While ecologists found natural selection an irrelevant component to explain relative SAD in forests [1,2], we found that the same simple neutral model fits SAD of genetic elements in genomes. Nature seems to play forest and genomes with same dice.

[1] Hubbell SP. The unified neutral theory of biodiversity and biogeography. Princeton University Press. (2001)

[2] Rosindell J, Hubbell SP, Etienne RS. The unified neutral theory of biodiversity and biogeography at age ten. Trends Ecol Evol 26 (2011) 340–348. 3.

[3] Brookfield JFY. The ecology of the genome - mobile DNA elements and their hosts. Nat Rev Genet 6 (2005) 128–136.

[4] Venner S, Feschotte C, Biémont C Dynamics of transposable elements: towards a community ecology of the genome. Trends Genet 25 (2009) 317–323

[5] Serra, F. & H. Dopazo. UNTB+. A Python package for large samples.

Phylogenetics and Evolution. Talk #3

Nested Phylogenetic Reconstruction: scalable resolution in large phylogenies.

Jaime Huerta-Cepas[1], Marina Marcet-Houben[1] and Toni Gabaldón[1]

1Bioinformatics and Genomics Programme. Centre for Genomic Regulation (CRG), UPF. Doctor Aiguader, 88. 08003 Barcelona (Spain).

A pressing challenge in phylogenetics is the need to cope with the massive production of complete genomic sequences, especially after recent technological developments. Problems that are particularly affected by the increasing flow of genomic data and that require continuous update are: i) the establishment of evolutionary relationships between species (the so-called Tree Of Life (TOL)) and ii) the study of the evolution of large, widespread super-families that evolved through complex patterns of duplications and losses. Current methods undertake phylogenetic recon- struction as a single step procedure, considering the whole set of species or homologous sequences as an indivisable dataset. We propose here the Nested Phylogenetic Reconstruction (NPR) approach, which can be used to recon- struct large phylogenies by means of an iterative strategy that provides scalable and sustained resolution in all tree nodes. In NPR, all internal nodes in a tree are re-evaluated to optimize and fine tune the parameters of phylogenetic inference.

To illustrate its use, we apply NPR to the reconstruction of a highly resolved eukaryotic TOL using a total of 216 fully sequenced species. Positive effects of the increased gene sampling at each iteration are clear, both in terms of balanced functional representation (Figure 1d), and of accuracy, as judged from the overall agreement with taxonomic classifications and established relationships (Figure 1a-c). The final topology is highly resolved, with all but 6 nodes in the tree receiving the highest statistical support. Agreement with established taxonomic divisions is remarkably high, considering our completely automated and uninformed approach. Indeed the final topology recovers the monophyly of 259 out of the 278 NCBI-based taxonomic groupings with two or more species in the tree, most inconsistencies corresponding to currently debated clades.

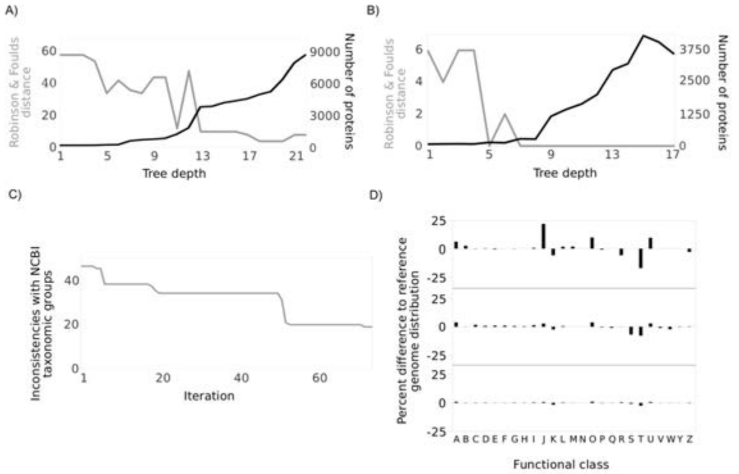


Fig1. TOL analyses: A-B) Grey lines represent topological distance between reference trees and the TOL (A-Chordates, B-Fungi). Black line represents the number of protein families used at each tree depth. C) Number of NCBI taxonomic groups not recovered at each iteration. D) Bars represent differences in percentage of protein families in different functional categories for proteins used in the TOL and the Human genome at three tree iterations: first node, base of chordates, and last iteration within primates.

Recent human evolution, continental differences in genes for complex disease and the common gene/common variant hypothesis.

Urko M. Marigorta[1] and Arcadi Navarro[1]

1. Institut de Biologia Evolutiva (UPF-CSIC) Barcelona (Spain).

The recent explosion of published GWAS (Genome Wide Association Studies) has permitted to identify hundreds of loci associated to human complex traits, particularly diseases. This discovery has widely improved our knowledge on disease susceptibility, quickly moving from a situation in which very few genes had been consistently replicated to the identification of a vast amount of new hits associated to diseases. For some time, there has been a strong bias towards individuals of European ancestry. This problem is being solved over the last few months with a flurry of GWAS performed upon non-Europeans that allowed us to tackle several important questions: are all risk variants equally replicated for all populations? To which extent do human populations share the genetic architecture of complex diseases?

To address these questions, we have selected 11 complex diseases for which two or more GWAS performed in Europeans and two or more GWAS performed in East Asians were available (as to August 2011). After careful selection of 109 disease-associated SNPs ($p < 5x10^{-7}$, or $5\ x10^{-8}$ if imputed) discovered in European GWAS, we have studied their degree of replication in all other GWAS. In total, we have gathered 210 replication attempts and have studied their patterns of replicability in each continent.

We report three main results. First, independently of the trait under study, SNPs that have been found to be significant in European GWAS present a high rate of replication ($p < 0.05$) in both European (≈ 90%) and East Asians (≈ 50%) populations. Second, we observe that the risk allele appears to be the same for > 90% of the replication attempts in East Asians, even when the associations are not significant. Moreover, the increase in disease risk (Odds Ratios) associated to the allele is highly correlated ($r = 0.75$, p-value$<10^{-15}$) between continents. Finally, the study of the patterns of differentiation in allele frequencies and linkage disequilibrium (LD) shows that genomic regions harbouring SNPs that consistently replicate between continents resemble more in their allele frequencies and LD patterns.

Our results show that (1) GWAS results are overwhelmingly real, i.e., non artifactual; (2) that common variants, and not at all rare variants, underlie the vast majority of significant results from GWAS; and (3) that the major contributors to genetic risk of most complex diseases, as unveiled by GWAS so far, are shared by populations from different continents. We discuss the implication of these findings for extant knowledge about the recent evolution of our species and for future advances in the study of the genetic architecture of complex disease.

Population genomics of 158 genomes of *Drosophila melanogaster*.

A. Barbadilla[1], S. Casillas[1], M. Barrón[1], D. Castellano[1] and M. Ràmia*

1. nstitut de Biotecnologia i Biomedicina and Department de Genètica i Microbiologia, Campus Universitat Autònoma de Barcelona, 08193, Cerdanyola (Barcelona), Spain.

The availability of 158 deep-coverage genomes from a natural population of Drosophila melanogaster represents an unprecedented opportunity to perform the most comprehensive nucleotide variation study done so far in any species. Here we provide a detailed genome-wide description of the nucleotide variation of this model species and by means of a battery of comparative methods for polymorphism and divergence data we try to answer questions of fundamental interest in population genomics such as: Which pattern or gradient follows genetic variation along the chromosomes? How these patterns correlate with structural regions or X vs. autosome arms? Which proportion of the coding and non-coding genome undergoes different selection regimes? How recombination rate modulate nucleotide variation and molecular evolution along the genome?

Poster presentations (Student Symposium)

Monday 23rd 9:00-18:00

Poster presentations (Student Symposium)

Monday 23rd 9:00-18:00

Phylogenomics supports microsporidia as the earliest branching group among sequenced fungi.

Salvador Capella-Gutiérrez, Marina Marcet-Houben & Toni Gabaldón

Bioinformatics and Genomics Programme. Centre for Genomic Regulation (CRG) and UPF. Barcelona, Spain.

Microsporida is one of the taxa that have experienced the most dramatic taxonomic reclassifications [1] since they were discovered about 150 years ago. Once thought to be among the earliest diverging eukaryotes, the fungal nature of this group of intracellular pathogens seems now widely accepted. However, the specific position of microsporidia within the fungal tree of life is still a matter of debate [2][3], with different studies supporting alternative scenarios. Due to the presence of accelerated rates of genome evolution, phylogenetic analyses involving microsporidia are prone to methodological artifacts such as long-branch attraction [4], especially when taxon sampling is limited. Here we exploit the recent availability of six complete microsporidian genomes to re-assess their phylogenetic position. We show that microsporidians have a similar low level of conservation of gene neighborhood with other groups of fungi, especially when controlling for the confounding effects of recent segmental duplications and differential shared gene content. A combined analysis of thousands of gene trees supports a topology in which microsporidia is sister group to all other sequenced fungi. Moreover, this topology received increased support when trees most likely to be affected by phylogenetic noise were discarded. This position of microsporidia was also strongly supported based on the combined analysis of 53 concatenated genes, and was robust to filters controlling for rate heterogeneity, compositional bias, long branch attraction and heterotachy. Moreover, we carried out simulations to test whether long-branch attraction artifact is enough to erode phylogenetic signal present in the alignment and attract microsporidia to a basal position in the tree independently they were placed during the simulations. Altogether, our data provide strong support for an earliest-branching position of microsporidia among sequenced fungi.

[1] Corradi, N., and PJ. Keeling. *Fungal Biology Reviews* **23** (2009) 1-8.
[2] Keeling, PJ. *PLoS Pathogens.* **5** (2009) e1000489.
[3] McLaughlin, D. J., D. S. Hibbett, F. Lutzoni, J. W. Spatafora, and R. Vilgalys. *Trends Microbiology*. **17** (2009) 488-497.
[4] Gribaldo, S., and H. Philippe. *Theor. Popul. Biol*. **61** (2002)391-408.

RNAseq-based identification and evolutionary analysis of phenol-degrading pathways in Candida yeasts

Leszek P. Pryszcz[1], Michaela Jakúbková[2], Gabriela Gérecová[2], Josef Nosek[2], and Toni Gabaldón[1]

[1] Department of Bioinformatics and Genomics, Centre for Genomic Regulation (CRG) and UPF, Dr Aiguader, 88, 08003 Barcelona, Spain
[2] Department of Biochemistry, Faculty of Natural Sciences, Comenius University, Mlynská dolina CH1, 842 15 Bratislava, Slovakia

Phenols and derived molecules are highly toxic to most organisms, yet some possess the ability to metabolize these compounds. For instance, several yeasts, such as the emerging fungal pathogen *Candida parapsilosis* are able to degrade various hydroxy derivatives of benzene and benzoic acid through the gentisate and 3-oxoapidate pathways. However, genetic control of these pathways in eukaryotes remains unknown. We analyzed transcriptomes (RNA-Seq) of *C. parapsilosis* cells cultivated on media with three carbon sources: glucose, 3-hydroxybenzoate and 4-hydroxybenzoate. Among genes highly up-regulated on phenolic compounds, we identified seven possibly coding for enzymes involved in these pathways. Recently, some of them were confirmed experimentally [1]. Here we identified remaining genes coding for the enzymes catalyzing reactions in both pathways and the putative transcription factors regulating switch of metabolism from glucose to hydroxybenzoates. Other up-regulated genes indicate the activation of stress-response pathways. In addition, we identified putative membrane pumps required for transport of 3-hydroxybenzoate and 4-hydroxybenzoate. Subsequently, the evolutionary history of the gentisate and 3-oxoapidate pathways in yeasts was analyzed. Complete pathways were found in 7 out of the 30 sequenced Saccharomycotina species. Among these, *Pichia stipitis* is known to metabolize hydroxybenzenes and hydroxybenzoates. We confirmed experimentally the ability to degrade 3-hydroxybenzoate and 4-hydroxybenzoate in another four species: *C. parapsilosis*, *C. orthopsilosis*, *C. metapsilosis*, and *C. subhashii* that are emerging human pathogens. This suggests that the gentisate and 3-oxoapidate pathways may be associated with pathogenesis and/or virulence. Despite the patchy phylogenetic distribution of phenol-degrading phenotype, we found no evidence of horizontal gene transfer. Rather, the pathways have evolved independently by functional divergence of existing enzymes.

[1] Z. Holesova et al. *Microbiology* **157** (2011) 2152-63.

Rapid and efficient contigs mapping and ordering with ICMapper

Hicham Benzekri[1], Antonio Muñoz-Merida[2] , Gonzalo Claros[1] and Oswaldo Trelles[2]

(1) The Andalusian Platform of Bioinformatics, University of Malaga, Spain

(2) Computer Architecture Department, University of Malaga, Spain >

To enhance efficiency in the gap closure phase of a genome project it is crucial to know which contigs are adjacent in the target genome. Rapid and efficient new tools are necessary to classify contigs from a draft assembly in order to get fully contiguous genome sequence. Therefore, with the advent of increasing numbers of finished genomes due to the incremental accessibility of new sequencing technologies and since there is an increasingly interest in sequencing closely related strains of existing finished genomes, it became possible to take advantage from this valuable information to guide whole genome assembly using bioinformatic pepeline.

ICMapper (Instant Contig Mapper) is a Perl program, intended as a tool to rapidly contiguate (*i.e.* align, order, orientate) assembled contigs based on a reference sequence. The underlying idea is that the remaining gaps between contigs for which no linkage information is present can subsequently be closed with direct PCR strategies. ICMapper program is based on summarizing resulting matches from pairwise alignment between draft genome and reference [1] and then order the corresponding contigs. When It has been compared with other implementations, specifically Oslay [1], Projector2 [2] and Mauve aligner [3] software using artificial contigs extracted from real finished genomes and using related genomes. ICMapper provides assembly results with accuracy rates up-to 50% better than those reported for Oslay and up-to 18% better against Projector2. Similar results are obtained when comparing with the last Mauve aligner. From the computational point of view, ICMapper get 20 times faster speed-up compared with state of the art software. Therefore, ICMapper outperform current software both, in the quality of contigs assembly and regarding computational resources, which make them a fast and efficient tool for completion of genome assembly processes.

Acknowledgements:

HB hold a grant from the Andalusian Platform of Bioinformatics, University of Malaga, Spain. AMM and OT are members of the GN5 node at the National Institute for Bioinformatics, Spain, a platform of ISCIII.

References:

[1] Richter,D.C. *et al*. (2007) OSLay: optimal syntenic layout of unfinished assemblies. *Bioinformatics*, 23, 1573–1579.

[2] van Hijum,S.A.F.T. et al. (2005) Projector 2: contig mapping for efficient gap-closure of prokaryotic genome sequence assemblies. Nucleic Acids Res., 33,W560–W566.

[3] Rissman, A.I. et al. (2009) Reordering contigs of draft genomes using the Mauve Aligner. Bioinformatics, 25, 2071-2073.

Modeling splicing from chromatin

Joao Curado[1,2], Hagen Tilgner[1,3], and Roderic Guigó[1]

[1] Centre for Genomic Regulation (CRG) and UPF, Barcelona, Spain
[2] GABBA, University of Porto, Porto Portugal
[3] Stanford University, Stanford, CA, U.S.A.

Although alternative splicing affects more than 90% of human genes and is proposed to be a primary driver of the evolution of phenotypic complexity in mammals, our ability to predict it is still low. There is increasing evidence that determinants outside of the RNA-sequence play a role in splicing. Introns are frequently removed from the nascent pre-mRNA transcript while RNA–Polymerase II continues transcription on the DNA template.

Using a statistical framework similar to that used to model transcription activity, we have developed predictive models to explore the relationship between levels of histone modifications in the vicinity of exons and the corresponding exon inclusion levels. The contribution of chromatin modifications to splicing is not expected to be dominant, and as a consequence, the predictive power of the exon inclusion models is smaller than that of the models of transcriptional activity. Nevertheless, a number of chromatin marks are identified consistently across cell lines as having a significant association with exon inclusion levels. Some with a strong positive association with exon inclusion, others with a clear negative one. Importantly, the models inferred in a particular cell type are, in general, quite accurate in the other cell lines, suggesting that the relationships between histone modifications and alternative splicing uncovered here appear to be general.

Unbound conformational ensembles can improve protein-protein docking predictions

Chiara Pallara, Juan Fernández-Recio

Joint BSC-IRB research programme in Computational Biology, Life Sciences Department, Barcelona Supercomputing Center, Spain

Structural prediction of protein-protein interactions is one of the major goals of computational biology, which will contribute to understand cellular processes at molecular level. Computer docking methods aim not only to predict the structure of a protein-protein complex, but also to understand the mechanism of protein association. Despite methodological advances, the experiment CAPRI (Critical Assessment of PRediction of Interactions)[1] has shown that the main bottleneck in docking is dealing with molecular flexibility and conformational changes. Indeed, our rigid-body docking method pyDock[2] had difficulties in cases with large conformational changes upon binding (unbound-bound Cα-atom ligand RMSD above 1.5 Å)[3].

There are different possible strategies to include flexibility in docking predictions. Assuming conformational selection mechanism, in which the unbound state can sample bound conformers, we have explored the use of precomputed unbound conformational ensembles in pyDock docking. To this aim, we first generated conformational ensembles for the unbound docking partners by using two different computational approaches, modelling minimization (MM) and molecular dynamics (MD).

Using all generated conformers in docking would involve an important combinatorial problem. The strategy here is to explore the validity of this approach by selecting different specific conformers from the ensemble: the ones with the best Cα and interface RMSD with respect to the bound state after superimposing onto the complex structure, the one with the best binding energy (based on our pyDock score) considering the complex orientation, and the structure built after clustering and selecting the most populated conformer per residue. Each of these unbound conformations for receptor and ligand were tested in pyDock docking.

We have tested this approach on a benchmark of 123 protein-protein complexes[4]. The results show a increase of pyDock performance when conformational movements upon binding are in the medium range. We will also explore whether the range of conformational flexibility of the ensembles could affect docking. The ultimate goal will be to develop a robust algorithm in order to include the most relevant conformers during the docking simulation. These results bring also interesting discussion on the protein-protein association mechanism (conformational selection vs. induced fit).

[1] Janin J Mol Biosyst. **6**(12) (2010) 2351–2362
[2] Cheng M.K.T et al. Proteins **68** (2007) 503-515
[3] Pons C. et al. Proteins **78** (2010) 95-108
[4] Hwang H. et al. Proteins **73** (2008) 705-709

Biblio-MetReS: A bibliomic network reconstruction application and server

Anabel Usié[1,2], Ivan Teixidó[1], Hiren Karathia[2], Rui Alves[2], Francesc Solsona[1] and Anabel Usié

[1]Departament d'Informàtica i Enginyeria Industrial, Universitat de Lleida, Av. Jaume II nº 69, 25001 Lleida, Spain.
[2]Departament de Ciències Mèdiques Bàsiques & IRBLleida, Universitat de Lleida, Montserrat Roig nº 2, 25008, Lleida, Spain.

Introduction: The study of gene and/or protein interactions in order to reconstruct molecular networks from automated analysis of the literature is one of the current goals of text-mining in biomedical research. Therefore, it is important to have methods that help reconstruct these networks structure. Some user-friendly tools such as iHOP [1,2] or String [3], perform this analysis using precompiled databases of abstracts from scientific papers. Other tools allow expert users to analyze the full content of a set of scientific documents. However, it would be useful to have a user-friendly tool that simultaneously analyzes the latest set of scientific documents available on line and reconstructs the set of genes referenced in those documents.

Results: We present such a tool, Biblio-MetReS. This tool does processing and text-mining of ascii, html and pdf documents found over Internet in order to reconstruct PPI networks. We benchmark the tool and compare its functioning and results against iHOP and String. All three tools create networks that are comparable, yet partially complementary because no two tools identify exactly the same set of PPIs in texts. Biblio-MetReS, provides more complete reconstructions than the other two tools, because it analyzes full text documents, as opposed to document abstracts [4,5].

Conclusions: Literature-based automated network reconstruction is still far from providing complete reconstructions of molecular networks. Nevertheless, its value as an auxiliary tool is already high and it will increase as standards or reporting biological entities and relationships become more widely accepted and enforced. Biblio-MetReS provides an easy environment for researchers to reconstruct their interest networks analyzing a set of scientific documents up to date. Biblio-MetReS can be downloaded from http://metres.udl.cat/

[1] R. Hoffman et al *Bioinformatics* **21** (2005) (Suppl 2): ii252-258.
[2] R. Hoffman et al. *Nature Genetics* **36** (2004) 664. (http://www.ihop-net.org)
[3] C. von Mering et al. nucleics Acids Res **35** (2007) D:358-362. (http://string-db.org/)
[4] PK Shah et al. BMC Bioinformatics **4** (2003) 20.
[5] J. Lin et al. *BMC Bioinformatics* **10** (2009) 46.

Poster presentations (odd numbers)

Day 2. **Odd numbered posters.** Tuesday 24th 9:00-19:00

Use of ChiP-Seq data for the design of a multiple promoter-alignment method

Ionas Erb[1], Juan Ramon González-Vallinas[2], Giovanni Bussotti[1], Enrique Blanco[3], Eduardo Eyras[2,4] and Cédric Notredame[1]

[1]Centre for Genomic Regulation (CRG) and UPF, Barcelona, Spain
[2]Pompeu Fabra University, Barcelona, Spain
[3]University of Barcelona, Barcelona, Spain
[4]Catalan Institution for Research and Advanced Studies (ICREA)

Alignments of regulatory regions are challenging for a number of reasons: the lack of structural constraints that can be exploited in the case of genes, a low information content of the sequence when considering independent nucleotides, and generally a lower identity of homologous regions. Another reason that has hampered progress is the lack of good benchmark sets, as up to now such alignments have commonly been tested on synthetically generated sequences whose value for evaluating performance on real-life data is limited. In this work we address these problems by testing a new method for multiple alignment of orthologous promoter regions on two newly developed benchmark sets. 'Pro-Coffee' uses a dinucleotide substitution matrix that was obtained from hundreds of functional binding sites in the TRANSFAC database. The use of dinucleotides enables us to take nearest-neighbor information into account during the alignment process. We compared its performance to a range of popular alignment algorithms (ClustalW, T-Coffee, Probcons, Muscle and Mafft). First, we designed a validation framework using several thousand families of orthologous promoters. This dataset was used to evaluate the accuracy for predicting true human orthologs among their paralogs. We used it to optimize the gap penalty scheme of the alignment methods to obtain maximum predictive performance. We found that whereas other methods achieve on average 73.5% accuracy, and 77.6% when trained on that same dataset, the figure goes up to 80.4% for Pro-Coffee. We then applied a novel validation procedure based on multi-species ChIP-seq data [1]. Trained and untrained methods were tested for their capacity to correctly align experimentally detected binding sites of CEBPA and HNF4a. Whereas the average number of correctly aligned sites for the two transcription factors is 284 for default methods and 316 for trained methods, Pro-Coffee achieves 331, 16.5% above the default average. We find a high correlation between a method's performance when classifying orthologs and its ability to correctly align proven binding sites. Not only has this interesting biological consequences, it also allows us to conclude that any method that is trained on the ortholog dataset will result in functionally more informative alignments. A characterization of alignments performing well in our tests reveals that they are neither short nor show they high column identities, contrary to what is commonly obtained from default methods.

[1] D. Schmidt et al. *Science* **328** (2010) 1036.

GRAPE RNAseq Analysis Pipeline Environment

David G. Knowles[*], Maik Röder, Roderic Guigó[†]

Barcelona 2011

Abstract

In this work we present a pipeline schema for the initial processing, analysis and management of RNAseq data, as well as the implementation of this pipeline using a workflow management system. This pipeline starts from the raw sequencing reads, or BAM alignments, a genome file and an annotation of the species from which the reads originate. It will run a set of quality control steps, align the reads if necessary and perform some standard analyses such as estimation of expression levels. The results are stored in a MySQL database that can be accessed using scripts or through a web application that allows for easy browsing of the results. GRAPE may be run locally or it can use a queuing system such as SGE to improve the speed and allow for parallelization of the different steps. New steps can be easily added, allowing it to be extended in order to fit specific needs.

[*] david.gonzalez@crg.cat

[†] roderic.guigo@crg.cat

A methylation caller to analyze DNA methylation levels

Daniel González-Peña[1], Osvaldo Graña[2], Florentino Fdez-Riverola[1], and David G. Pisano[2]

[1]*ESEI: Escuela Superior de Ingeniería Informática, University of Vigo, Edificio Politécnico, Campus Universitario As Lagoas s/n, 32004, Ourense, Spain.*
[2]*Bioinformatics Unit, Structural Biology and BioComputing Programme. Spanish National Cancer Research Centre (CNIO). C/ Melchor Fernández Almagro 3, 28029, Madrid, Spain.*

Epigenetics investigates changes in gene expression by other factors different to the own DNA sequence [1]. These changes in gene expression are the answer to chemical modifications of the DNA and its associated proteins. DNA methylation is a chemical process where a methyl group is added to cytosine nucleotides. It was known that the methylated cytosines were followed by a guanine [2][3][4], but recent studies have shown that some cytosines that are also methylated are positioned in a free guanine context [5][6][7].
The new technologies for DNA sequencing had opened great possibilities to the discovery of new DNA methylation patterns and its implications. Altering DNA methylation can contribute to epigenetic modifications that can produce abnormal activation or silencing of genes. Such alterations have been associated with syndromes involving chromosomal instability, mental retardation and cancer [8][9][10].
From the bionformatics point of view, new software is required to analyze DNA methylation levels from next generation sequencing experiments. Several tools are available nowadays [11][12][13][14][15], but they still lack functionalities in the analysis that are demanded by molecular biologists. The software package that we present here performs a full DNA methylaton level analysis for all of the cytosines that are present in one of the three known methylation contexts. Having previously tested the efficiency of the protocol used here [16], we added statistical significance to the results in a similar way to the work done by Lister et al. Furthermore, this software calculates global methylation levels by genomic context and DNA strand and allows users to annotate the methylated cytosines against any set of known annotations that are introduced as bed files.

[1] Holliday R. *Developmental genetics* 15 (1994) 453-457.
[2] Jones P.A. and Baylin, S.B. *Nat. Rev. Genet.* 3 (2002) 415-428.
[3] Robertson K.D. *Oncogene* 21 (2002) 5361-5379.
[4] Egger G. et al. *Nature* 429 (2004) 457-463.
[5] Lister R. et al. *Cell* 133 (2008) 523-536.
[6] Cokus S.J. et al. *Nature* 452 (2008) 215-219.
[7] Lister R. et al. *Nature* 462 (2009) 315-322.
[8] Egger G et al. Nature 429 (2004) 457-463.
[9] Martin D.I.K et al. *Epigenetics* 7 (2011) 843-848.
[10] Esteller M. *Genes Cancer* 2 (2011) 6:604-606.
[11] Xi Y. and Wei L. *BMC Bioinformatics* 10 (2009) 232.
[12] Harris E.Y. et al. Bioinformatics 26 (2009) 572-573.
[13] Chen P.Y. et al. *BMC Bioinformatics* 11 (2010) 203.
[14] Krueger F, Andrews SR. *Bioinformatics* 27 (2011) 11:1571-1572.
[15] Pedersen B. et al. *Bioinformatics* 27 (2011) 17:2435-2436.
[16] González-Peña et al. *5th International Conference on Practical Applications of Computational Biology and Bioinformatics 6-8th. Springer* (2011) 87-91.

Integrative biology for the study of Malignant Peripheral Nerve Sheath Tumors

Bernat Gel[1], Ernest Terribas[1], Jaume Mercadé[1], Conxi Lazaro[2], Eduard Serra[1]

[1] *Grup de Cancer Hereditary, Institut de Medicina Predictiva i Personalitzada del Càncer (IMPPC), Badalona, Spain*
[2] *Laboratori de Recerca Translacional, Institut Català d'Oncologia; Institut d'Investigació Biomèdica de Bellvitge (IDIBELL), L'Hospitalet de Llobregat, Barcelona, Spain.*

One of the major clinical complications of Neurofibromatosis type 1 (NF1) patients is the development of different tumor types. Peripheral nervous system, (PNS) is particularly affected, with the development of multiple dermal neurofibromas (dNFs), plexiform neurofibromas (pNFs) and malignant peripheral nerve sheath tumors (MPNSTs). Approximately 8–13% of NF1 patients develop MPNSTs, which are the leading cause of NF1-related mortality. Due to its invasive growth, propensity to metastasize, and limited sensitivity to chemotherapy and radiation, MPNST has a poor prognosis, with a 5-year survival rate of NF1 patients with MPNST of 21%.

We analyzed different expression profiles obtained for dNFs, pNFs and MPNSTs, and derived cell lines, to better understand MPNST development and malignant progression. These expression profiles were used to identify groups of frequently up-regulated or down-regulated genes that were clustering together in specific genomic regions, what are called transcriptional imbalances (TIs).

SNP-array analysis in 5 MPNST-derived cell lines and 14 MPNSTs showed that MPNSTs contain a genome with a high degree of aneuploidy and many somatic copy-number alterations (SCNA). Furthermore, MPNSTs present a high frequency of LOH events including large copy-neutral LOH regions.

In addition, the complete exomes of 3 MPNST-derived cell line and 2 MPNST xenografted tumors have been sequenced to start cataloguing all possible point mutations and small insertions/deletions in coding genes that occur in this type of tumors.

Finally, epigenetic data including CpG island methylation levels and histone modification marks are being collected to generate epigenetic profiles of MPNST tumors and derived cell lines. This information will be used to detect the presence of Long-Range Epigenetic Silencing (LRES), the phenomenon by which chromatin remodeling suppresses the expression of genes placed contiguous in specific genomic locations.

An integrative analysis of all this data along with previously published data including cancer genes, MPNST copy-number profiles obtained from aCGH analysis and MPNST expression data, will allow us to gain new insights on the biology of MPNSTs and the mechanisms driving its development and progression.

Genome Maps, a new genomic HTML5 web based graphical analysis tool

Ignacio Medina[1,3], Alejandro de María[1], Francisco Salavert[2] and Joaquín Dopazo[1,2,3]

[1]*Department of Bioinformatics and Genomics, Centro de Investigación Príncipe Felipe (CIPF), Valencia, Spain*
[2]*CIBER de Enfermedades Raras (CIBERER), Valencia, Spain*
[3]*Functional Genomics Node (INB), CIPF, Valencia, Spain*

Last years data generated by high-throughput technologies like microarrays and next generation sequencing have been flooding biological repositories and databases with relevant information. This information covers regulatory regions, functional and genomic variation between others. As more new data is being obtained we need new tools that allow researchers to visualize all the data, and integrate and analyze experiments in a graphical manner.

We have developed the first web based genomic graphical analysis tool completely based on the new web standard HTML5, by using this new standard we have been able to visualize genomes in SVG directly in web browser, we have also developed a complete set of widgets that allow researchers to work with their expression or genotype data.

Between the most relevant features we can find:

- 100% web based in *HTML5* so there is no need of software or plugin installation
- Most relevant biological information has been integrated in the web application as these data is exported in *JSON* by our *RESTful* web services from our servers, these data includes: genes, transcripts, exons, cytobands, *SNPs*, mutations, miRNAs targets, regulatory regions, *CTFC,* ...
- DAS servers are also easily integrable as a DAS client parser has been developed
- Some analysis have been integrated to be able to work with expression and genotype data
- Researchers can save their data and session as they can be logged in our servers and 1 GB of free data is given

IDENTIFYING GENE NETWORKS THAT MEDIATE THE MYOCARDIAL PHENOTYPE OF CD36 DEFICIENCY: A GLOBAL GENE EXPRESSION ANALYSIS IN THREE MICE MODELS

I. SABOUNI (1), I. MEZIANE (1), N.A. ABUMRAD (2), A. MOUSSA (3),
Y. CHERRAH (1) and A. IBRAHIMI (1)
(1) Laboratory of Pharmacology and Toxicology, Medicine and Pharmacy School,
Mohammed V Souissi University, Rabat, Morocco
(2) Department of Medicine, Washington University School of Medicine,
Center for Human Nutrition, St Louis, Missouri, USA
(3) Ecole Nationale des Sciences Appliquées, Abdelmalek Essaadi University, Tanger,
Morocco

We have previously shown that CD36 is a membrane protein that facilitates long chain fatty acid (FA) transport by muscle tissues. We also documented the significant impact of muscle CD36 expression on heart function, skeletal muscle insulin sensitivity as well as on overall metabolism. To search for a comprehensive set of genes that are differentially regulated by CD36 expression in the heart, we conducted DNA microarray with tissues from CD36 Knockout (KO) versus wild type (WT) mice. To validate specificity of observed changes with respect to CD36 expression, we also used tissues from a CD36 null mouse rescued for CD36 in heart and skeletal muscle (GR).

Absence of CD36 led to down-regulation of the expression of three groups of genes involved in pathways for FA metabolism, angiogenesis/apoptosis and structure. These data are consistent with the fact that the CD36 protein recognizes a number of ligands in addition to FA, including the anti-angiogenic factor Thrombospondin 1 (Tsp1). Furthermore, compensatory changes were also identified with over expression of sets of genes linked to these same pathways. Changes were largely reversed by rescue of CD36 in GR mice.

In conclusion, the study supports the interpretation that rescuing expression of CD36 protein in hearts of CD36 null mice reverses the major abnormalities linked to CD36 deficiency in large part through regulation of gene expression. The findings also identify specific pathways as possibly underlying the phenotypic abnormalities. The findings will serve to build a bioinformatics model to derive insight regarding involvement of specific pathways and gene networks in the phenotypic effects of CD36 deficiency.

Keywords : CD36; Microarrays;metabolism, angiogenesis/apoptosis,gene networks

DIGEN-1K. Experimental development of a System for Genetic Diagnostics and Pathogen Identification throughout Genomic Sequencing (NGS).

Jan-Jaap Wesselink[1,2], Silvia Lusa[1,2], Santiago de la Peña[1,2] and Paulino Gómez-Puertas[1].

[1] *Molecular Modelling Group. Centro de Biología Molecular "Severo Ochoa" (CSIC-UAM), C/ Nicolás Cabrera, 1, Cantoblanco, 28049 Madrid, Spain.*
[2] *Biomol-Informatics, Parque Científico de Madrid, C/ Faraday, 7, Cantoblanco, 28049 Madrid, Spain.*

We present the DIGEN-1K initiative: "Experimental development of a System for Genetic Diagnostics and Pathogen Identification throughout Genomic Sequencing (NGS)", a INNPACTO research project lead by **Biomol-Informatics S.L.** in cooperation with three main public research organizations: **Parque Científico de Madrid**, **Instituto de Investigación Sanitaria del Hospital Universitario La Paz (IdiPAZ)** and **Centro de Biología Molecular "Severo Ochoa" (CBMSO, CSIC-UAM).**

The project is structured in three parallel workpackages:

WP1.- Optimization of the system for NGS data analysis of human exomes, including detection of SNV and prediction of the effects protein functionality (Centered in cancer and Rare Diseases).

WP2.- Optimization of Human exome sequencing processes through exome capture and NGS techniques.

WP3.- Pathogen identification through genome sequencing using NGS.

The pipeline for data analysis optimization of human exome sequences includes: Base calling, sample demultiplexing and quality assignment of reads, using the standard "Illumina Data Analysis Pipeline" software and alignment of reads to reference human genome (hg19, UCSC assembly, Feb. 2009), applying the Burrows-Wheeler (BWA) alignment tool [1]. Only sequences with an optimal quality over a specified threshold will be used to localize Single Nucleotide Variations (SNVs) as well as insertions and deletions (indels). The "Genome Analysis Toolkit (GATK)" [2] will be used to re-align and recalibrate the reads and subsequent identification of SNVs and indels in the re-aligned sequences. The list of reported SNVs will be extracted from the last version of the dbSNP database (National Center for Biotechnology Information, NCBI) [3], as well as from the 1000 Genomes Project Consortium [4]. Prediction of synonymous and non-synonymous substitutions will be done using the SIFT algorithm [5] and the Ensembl Variant Effect Predictor [6].

The pipeline has been automated using independent modules, currently allowing the simultaneous analysis of a large number of samples (300-500 human exomes/month). In addition, the automated pipeline uses a variety of quality filters (fastx-toolkit [7], prinseq [8], picard [9], fastqc [10]), incrementing not only the velocity but also the quality of the automated analysis. Finally, a multi-purpose, user-friendly web-based platform is being implemented to share data with the researchers, including quality tables, SNVs identification, genome browsing, etc.
An example of human exome analyzed through the automated pipeline, as well as full set of technical details, results, predictions, etc. will be accessible at "www.exome-ngs.com".

[1] Li & Durbin. *Bioinformatics* 25 (2009) 1754–1760
[2] McKenna et al. *Genome Res.* 20 (2010) 1297–1303
[3] http://www.ncbi.nlm.nih.gov/SNP/
[4] 1000 Genomes Project Consortium. *Nature* 467 (2011) 1061–1073
[5] Kumar & Henikoff. *Nat Protoc* 4 (2009) 1073–1081
[6] McLaren et al. *BMC Bioinformatics* 26 (2010) 2069-2070
[7] http://hannonlab.cshl.edu/fastx_toolkit/
[8] Schmieder & Edwards. Bioinformatics 27 (2009) 863—864
[9] http://picard.sourceforge.net
[10] http://www.bioinformatics.bbsrc.ac.uk/projects/fastqc/

RNA-seq time course analysis of a cell transdifferentiation system.

A. Conesa[1], S. Tarazona[1], R. Lomas[1], M.J. Nueda[2], J. Carcel[3],
C. García[3], C. Simón[3], and Pardo M.A.[1]

[1]Group of Genomics of Gene Expression. Bioinformatics and Genomics Department. Centro de Investigaciones Príncipe Felipe. Valencia. (mpardo@cipf.es)
[2] Department of Statistics and Operation Research. Universidad de Alicante.
[3]Regenerative Medicine Department. Centro de Investigaciones Príncipe Felipe. Valencia.

A new generation of sequencing technologies (NGS) has provided new opportunities for high-throughput functional genomic research [1]. To date, these technologies have been applied in a variety of contexts, including whole-genome sequencing, targeted resequencing and noncoding RNA expression profiling. One of the applications of these new deep-sequencing technologies is the study of dynamic transcriptomes in an approach termed RNA-seq [2].

NGS data consist of a list of short sequences, in the case of RNA-seq they come from the RNA molecules contained in the cell. Specific tools that combined computational and statistical algorithms are required to extract the biological information from the data.

Recently studies have characterized the transdifferentiation of human cancer cells into adipocyte-like cells, induced by the presence of specific unsaturated fatty acids [3]. In this project, a RNA-seq approach has been applied to the study of the transdifferentiation of melanoma MALME-3M. A set of data from different states of this process (time from incubation with unsaturated fatty acids: 0 h, 3 h, 7 h, 11 h, 15 h and 24 h) is available. The bioinformatic analysis of this data would allow us the quantification of the changing expression levels of each transcript and the gene regulatory changes that take place in this process. In addition, this study has been extended to two gene regulatory mechanisms related to cancer: antisense expression and RNA editing. For this purpose, it is necessary to design new strategies and tools. All the information obtained could be useful to determine which are the mechanism that drives this transdifferentiation.

[1] M. Metzker. *Nat Rev Genet* **11** (2010) 31.
[2] S. Marguerat et al. *Cell Mol Life Sci* **67** (2010) 569.
[3] A. Ruiz-Vela et al. *Stem Cell Rev and Rep* (2011).

Extracting high-quality methylation maps from HTS reads

Michael Hackenberg, José L. Oliver, Antonio Rueda, Francisco Dios, and Guillermo Barturen

Dpto. de Genética, Facultad de Ciencias, Universidad de Granada, Campus de Fuentenueva s/n, 18071-Granada, Spain & Lab. de Bioinformática, Centro de Investigación Biomédica, Campus de la Salud, Avda. del Conocimiento s/n, 18100-Granada, Spain

Whole genome methylation profiling at single cytosine resolution is now feasible due to the advent of high-throughput sequencing techniques, together with bisulfite treatment of the DNA. Several tools [1-3] have been proposed to cope with the challenge of converting the enormous amounts of sequencing data into high quality methylation maps. Nevertheless, the focus so far has been on speed and not on the quality control of the inferred methylation levels. Here, we present a pipeline that focus on the resulting methylation maps quality, without forgetting the hardware requirements, versatility and speed. Furthermore, we present a comprehensive analysis of the main parameters affecting the mapping quality and cytosine coverage, such as seed length or number of mismatches within the seed; we present as well a new approach to select the best alignments of the reads. NGSmethPipe [4] is the first program implementing the detection of SNVs (single nucleotide variants, [5]), bisulfite failures [6] and a strict control of sequencing errors, together with an improved mapping strategy based on the Bowtie aligner [7]. Other important technical features like full multi-threading and scalable memory requirement are also implemented.
Availability: http://bioinfo2.ugr.es/NGSmethPipe/.

[1] Pedersen, B et al. *Bioinformatics (Oxford, England)* **27** (2011) 2435-2436
[2] Krueger, F et al. *Bioinformatics (Oxford, England)* **27** (2011) 1571-1572
[3] Chen, PY et al. *BMC Bioinformatics* **11** (2011) 203
[4] Barturen, G. et al. (submitted)
[5] Weisenberger, DJ et al. *Nucleic acids research* **33** (2005) 6823-6836
[6] Lister, R et al. *Nature* **462** (2009) 315-322
[7] Langmead, B et al. *Genome biology* **10** (2009) R25

A new view on the functional role of alternative splicing: the impact of stochastic fluctuations on cell-to-cell variations

Santi Béjar[1], Jordi Morata[1], Casandra Riera[1] and Xavier de la Cruz[1,2]

[1] IBMB-CSIC, Computational Biology and Bioinformatics, 08028 Barcelona, SPAIN
[2]Institució Catalana de Recerca i Estudis Avançats (ICREA)

BACKGROUND

In recent years a growing amount of both experimental and computational works have shown that gene expression is stochastic, that is, gene product levels vary within tissue cell populations. This stochastic behavior is not an irrelevant physico-chemical effect; it has been shown to play a key role in biological processes, including differentiation and development [1]. Until now studies of stochastic gene expression have been focused on the conventional "one gene - one protein" scheme; this situation has drastically changed with the recent publication of a seminal work by Waks et al. [2], who have provided the first study of AS cell-to-cell variability. The results of this study open a completely new research pathway in the field of functional alternative splicing (AS) and in the interpretation of its experimental results. Among the opened questions are: what is the real functional impact of AS differences between tissues? Is the tissue level interpretation of some AS experiments consistent with the underlying cell population's picture? Etc. In this presentation I will describe our computational model for the stochastic modeling of alternative splicing, the main results obtained and will show how they can be used to address some of these questions.

RESULTS

For several model genes, we have systematically explored the impact of AS on the mRNA and protein distribution over cells. From a technical point of view we find that apart from changes in raw amount, AS has little impact on the stochastic behavior of the main isoform. More interestingly, we observe that even for moderate alternative splicing (average % of alternative isoform between 20% and 30%), there is always a non-negligible number of cells that retain their main isoform normal levels. We also find that the distribution of alternative isoform levels can display drastically different behaviors depending on the mRNA or protein properties. This last result casts a word of caution on the interpretation of some experimental approaches based on the use of tissue samples (averaged cell populations) to study AS.

CONCLUSIONS

We have used a computational model to study AS within the framework of stochastic gene expression. Our main results provide a simple, but clarifying view, of cell-to-cell variations in AS, and can be used to provide new hints on the contribution of AS to tissue differentiation. They also shed light on the interpretation of standard experimental approaches to the study of AS.

REFERENCES

[1] Raj, A. and van Oudenaarden, A. *Cell* **135** (2008) 216.
[2] Waks, Z. et al., *Mol Sys Biol* **7** (2011) 506.

Modelling and Simulating Generic RNA-Seq Experiments with the Flux Simulator

Thasso Griebel[1], Benedikt Zacher[1,4], Paolo Ribeca[2], Emanuele Raineri[3], Vincent Lacroix[5,6], Roderic Guigó[6], Michael Sammeth[1,6]

[1]Functional Bioinformatics Group, [2]Algorithm Development Group, [3]Statistical Genomics Group, Centro Nacional de Análysis Genómico (CNAG), 08028 Barcelona, Spain

[4]Computational Biology and Regulatory Networks Group, Gene Center Munich, 81377 Munich, Germany

[5]Biométrie et Biologie Évolutive, Université Lyon 1, 69622 Villeurbann, France

[6]Bioinformatics and Genomics Program, Centro de Regulación Genómica, 08003 Barcelona, Spain

High-throughput sequencing of cDNA libraries constructed from entire cellular RNA complements (RNA-Seq) naturally provides a digital quantitative measurement for every expressed RNA molecule. However, nature, impact and mutual interference of biases in different experimental setups are still poorly understood, mostly due to the lack of data from intermediate protocol steps. We analyzed multiple RNA-Seq experiments across different sample preparation protocols and sequencing platforms, and we investigated common—and currently indispensable—technical processes that influence abundance and distribution of sequenced reads: reverse transcription, fragmentation, final library preparation, and the sequencing step. For each of those steps, we developed universally applicable models, which can be parameterized by empirical attributes of corresponding sample preparations. Our models are implemented in a computer simulation pipeline called the Flux Simulator, and we show by experimental evidence that *in silico* RNA-Seq provides insights about precursors in an experiment that determine different read distributions observed in the outcome of specific control sequences. We further demonstrate enhancing and compensatory effects of reported biases, and we identify additional sources of systematic influences by the currently applied RNA hydrolysis step. Finally, we reproduce by different combinations of our models the experimental outcome of substantially different experimental setups, covering the major protocols proposed for RNA-Seq so far.

aGEMv3.1: Does mouse look like human (at the expression level)?

Segura J[1], Jiménez-Lozano N[2,3], Macías JR[3,4], Vega J[3,4], Carazo JM[2,3,4]

[1]Leeds Institute of Molecular Medicine, Section of Experimental Therapeutics, Welcome Trust Brenner Building, St. James's University Hospital, Leeds LS9 7TF, United Kingdom.
[2]GN7 of the National Institute for Bioinformatics (INB)
[3]BCU of the National Center for Biotechnology (CNB-CSIC). Darwin, 3. 28049 Madrid, Spain.
[4]Instruct Image Processing Center.

Information concerning the gene expression pattern in four dimensions (species, genes, anatomy and developmental stage) is crucial for unraveling the roles of genes through time. There are a variety of anatomical gene expression data bases, but extracting information from them can be hampered by their diversity and heterogeneity.

aGEM 3.1 (anatomic Gene Expression Mapping) addresses the issues of diversity and heterogeneity of anatomical gene expression data bases by integrating six mouse gene expression resources (EMAGE, GXD, GENSAT, Allen Brain Atlas data base, EUREXPRESS and BioGPS) and three human gene expression data bases (HUDSEN, Human Protein Atlas and BioGPS). Furthermore, aGEM 3.1 provides new cross analysis tools to bridge these resources.

aGEM 3.1 (http://agem.cnb.csic.es) can be queried using gene and anatomical structure. The first type of query can be carried out by using the ENSEMBL identifier, common gene symbol, MGI identifier or UniProt accession number. Querying by anatomical structure is simplified by displaying the terms from the chosen developmental stage (Carnegie Stage for human or Theiler Stage for mouse) in hierarchical trees.

By integrating the KEGG pathways data base in aGEMv3.1, the user can query using a set of genes involved in a given process. This utility is very useful as it allows gene expression differences and similarities to be compared in physiological and disease states. Emphasizing the disease state, information from OMIM (human diseases) and MTB (tumors in mice) data bases is displayed associated with genes in the results interface, when available.

Output information is presented in a friendly format, allowing the user to display expression maps and correlation matrices for a gene or structure during development. An in depth study of a specific developmental stage is also possible using heat maps that relate gene expression with anatomical components. Currently, aGEM provides a friendly and intuitive display of gene expression information for more than 1,450,000 and 278,000 gene-structure pairs for mouse and human, respectively.

[1] Jiménez-Lozano N et al. *Bioinformatics* (2011); doi: 10.1093/bioinformatics/btr639.

Gitools: Analysis and Visualisation of Genomic Data Using Interactive Heat-Maps

Christian Perez-Llamas[1] and Nuria Lopez-Bigas[1,2]

1 Research Unit on Biomedical Informatics, Department of Experimental and Health Sciences, University Pompeu Fabra, Barcelona, Spain.

Intuitive visualization of data and results is very important in genomics, especially when many conditions are to be analyzed and compared. Heat-maps have proven very useful for the representation of biological data. Here we present Gitools (http://www.gitools.org), an open-source tool to perform analyses and visualize data and results as interactive heat-maps. Gitools contains data import systems from several sources (i.e. IntOGen, Biomart, KEGG, Gene Ontology), which facilitate the integration of novel data with previous knowledge.

Differential expression in RNA-seq: A matter of depth

Sonia Tarazona[1,2], Fernando García-Alcalde[1], Joaquín Dopazo[1], Alberto Ferrer[2], and Ana Conesa[1]

1 Bioinformatics and Genomics Department, Centro de Investigación Príncipe Felipe, Valencia, Spain. 2 Department of Applied Statistics, Operations Research and Quality, Universidad Politécnica de Valencia, Valencia, Spain.

Next-generation sequencing (NGS) technologies are revolutionizing genome research, and in particular, their application to transcriptomics (RNA-seq) is increasingly being used for gene expression profiling as a replacement for microarrays. However, the properties of RNA-seq data have not been yet fully established, and additional research is needed for understanding how these data respond to differential expression analysis. In this work, we set out to gain insights into the characteristics of RNA-seq data analysis by studying an important parameter of this technology: the sequencing depth. We have analyzed how sequencing depth affects the detection of transcripts and their identification as differentially expressed, looking at aspects such as transcript biotype, length, expression level, and fold-change. We have evaluated different algorithms available for the analysis of RNA-seq and proposed a novel approach -NOISeq- that differs from existing methods in that it is data-adaptive and nonparametric. Our results reveal that most existing methodologies suffer from a strong dependency on sequencing depth for their differential expression calls and that this results in a considerable number of false positives that increases as the number of reads grows. In contrast, our proposed method models the noise distribution from the actual data, can therefore better adapt to the size of the data set, and is more effective in controlling the rate of false discoveries. This work discusses the true potential of RNA-seq for studying regulation at low expression ranges, the noise within RNA-seq data, and the issue of replication.

Sequence shortening in the rodent ancestor

Steve Laurie[1], Macarena Toll-Riera[1], Núria Radó-Trilla[1], M.Mar Albà[1,2]

1 Evolutionary Genomics Group, Pompeu Fabra University (UPF) and Municipal Institute of Medical Research (FIMIM), Barcelona, Spain. 2 Catalan Institution for Research and Advanced Studies (ICREA), Barcelona, Spain.

Insertions and deletions (indels), together with nucleotide substitutions, are major drivers of sequence evolution. An excess of deletions over insertions in genomic sequences – the so-called deletional bias – has been reported in a wide range of species, including mammals. However, this bias has not been found in the coding sequences of some mammalian species, such as human and mouse. To determine the strength of the deletional bias in mammals and the influence of mutation and selection, we have quantified indels in both neutrally evolving non-coding sequences and protein-coding sequences, in six mammalian branches – human, macaque, ancestral primate, mouse, rat and ancestral rodent. The results obtained with an improved algorithm for the placement of insertions in multiple alignments, Prank+F, indicate that contrary to previous results, the only mammalian branch with a strong deletional bias is the rodent ancestral branch. We estimate that such a bias has resulted in approximately 2.5% sequence loss of mammalian syntenic region in the ancestor of mouse and rat. Further, a comparison of coding and non-coding sequences shows that negative selection is acting more strongly against mutations generating amino acid insertions than against mutations resulting in amino acid deletions. The strength of selection against indels is found to be higher in the rodent branches than in the primate branches, consistent with the larger effective population sizes of the rodents.

Evaluation of genomic high-throughput sequencing data generated on Illumina HiSeq and Genome Analyzer systems

André E Minoche[1,2], Juliane C Dohm[1,2] and Heinz Himmelbauer[2]

1 Max Planck Institute for molecular Genetics, Ihnestr. 63-73, 14195 Berlin, Germany. 2 Centre for Genomic Regulation (CRG) and UPF, C. Dr. Aiguader 88, 08003 Barcelona, Spain.

Background: The generation and analysis of high-throughput sequencing data is becoming a major component of many studies in molecular biology and medical research. Illumina's Genome Analyzer (GA) and HiSeq instruments are currently the most widely used sequencing devices. Here, we comprehensively evaluate properties of genomic HiSeq and GAIIx data derived from two plant genomes and one virus, with read lengths of 95-150 bases.

Results: We provide quantifications and evidence for GC bias, error rates, error sequence context, effects of quality filtering, and the reliability of quality values. By combining different filtering criteria we reduced error rates 7-fold at the expense of discarding 12.5% of alignable bases. While overall error rates are low in HiSeq data we observed regions of accumulated wrong base calls. Only 3% of all error positions accounted for 24.7% of all substitution errors. Analyzing the forward and reverse strand separately revealed error rates of up to 18.7%. Insertions and deletions occurred at very low rates on average but increased to up to 2% in homopolymers. A positive correlation between read coverage and GC content was found depending on the GC content range.

Conclusions: The errors and biases we report have implications on the use and the interpretation of Illumina sequencing data. GAIIx and HiSeq data sets show slightly different error profiles. Quality filtering is essential to minimize downstream analysis artifacts. Supporting previous recommendations, the strand-specificity provides a criterion to distinguish sequencing errors from low abundance polymorphisms.

The role of Gln 61 in HRas GTP hydrolysis: a Quantum Mechanics/ Molecular Mechanics (QM/MM) study

Fernando Martín-García[1,2], Jesús Ignacio Mendieta-Moreno[1,2], Eduardo López-Viñas[1,2], Paulino Gómez-Puertas[1] and Jesús Mendieta[1,2]

1 Molecular Modelling Group. Centro de Biología Molecular "Severo Ochoa" (CSIC-UAM), C/ Nicolás Cabrera, 1, Cantoblanco, 28049 Madrid, Spain. 2 Biomol-Informatics, Parque Científico de Madrid, C/ Faraday, 7, Cantoblanco, 28049 Madrid, Spain.

Activation of the water molecule involved in GTP hydrolysis within the HRas–RasGAP system is analyzed using a tailored approach based on hybrid *QM/MM* simulation. A new path emerges: transfer of a proton from the attacking water molecule to a second water molecule; then a different proton is transferred from this second water molecule to the GTP. Gln 61 will stabilize the transient OH- and H3O+ molecules thus generated. This newly proposed mechanism was generated by using for the first time the entire HRas-RasGAP protein complex in a *QM/MM* simulation context. It also offers a rationale explanation for previous experimental results as the decrease of GTPase rate found in the HRas Q61A mutant and the increase exhibited by the HRas Q61E mutant.

Computer prediction of protein-RNA complex structure by docking and statistical potentials

Laura Pérez-Cano[1] and Juan Fernández-Recio[1]

[1]*Life Sciences Department, Barcelona Supercomputing Center (BSC), Jordi Girona 29, Barcelona, 08034, Spain*

Despite the importance of protein-RNA interactions in the cellular context, the number of available protein-RNA complex structures is still much lower than that of other biomolecules. As a consequence, an increasing number of computational studies are nowadays addressed towards structural prediction of protein-RNA complexes. In this context, we extracted residue propensities for the prediction of RNA-binding sites, which were implemented in our algorithm OPRA (Optimal Protein-RNA Area) [1]. This method predicted RNA-binding areas on proteins with 80 % PPV. This method has helped to identify the RNA binding region on spTranslin protein [2]. We have also extracted pairwise residue-ribonucleotide statistical potentials for protein-RNA docking. The approach is based on FTDock generation of rigid-body docking poses with a further fast scoring based on the pairwise statistical potentials [3]. This protocol was successfully applied for the structural prediction of a blind test case in CAPRI competition [4] and is now being useful to determine the structure of the human Translin-RNA complex. In order to help to further improve docking protocols and find better parameters for protein-RNA predictions, we have collected a non-redundant protein-RNA docking benchmark. It is composed of 106 test cases comprising 9 unbound-unbound, 8 model-unbound, 5 unbound-model, 3 model-model, 62 unbound-bound and 19 model-bound protein-RNA cases. The version 1.0 of the benchmark will be freely available on Internet to foster the development of protein-RNA docking algorithms and to contribute to the better understanding and prediction of protein-RNA interactions

[1] Pérez-Cano et al. *Proteins* **78** (2010)
[2] Eliahoo et al. *Nucleic Acids Res* **38** (2010)
[3] Pérez-Cano et al. *Pac Symp Biocomput* **2010** (2010)
[4] Pons et al. *Proteins* **78** (2010)

A Combined Information Retrieval and Text mining Classification framework for MHC Peptide Binding Prediction

Fadi Chakik[1], Ahmad Shahin[1], and Walid Moudani[1]

[1] *LaMA Group, Lebanese University, email: {fchakik, ashahin, wmoudani}@ul.edu.lb*

A key step in the development of an adaptive immune response to vaccines is the binding of peptides to molecules of the Major Histocompatibility Complex (MHC) for presentation to T lymphocytes, which are thereby activated. Several algorithms have been proposed for such binding predictions, but these are limited to a small number of MHC molecules and have imperfect prediction power. We are undertaking an exploration of the power gained by taking advantage of a natural representation of the protein sequence amino acid in terms of their composition, structural and a series of associated physicochemical properties [1] [2] to form a representative descriptor vectors. We are proposing to use a combined algorithm derived from information retrieval theory and text mining algorithm [3] for preprocessing and generating the descriptor vectors of the amino acids sequences before feeding them into a well known statistical classifiers [4] [5] for binding prediction. This algorithm will leads to substantially higher values for our evaluation criteria (Area Under ROC Curve) which means that misclassification errors is reaching lower rates.

[1] J. R. Bock, and D. A. Gough, "Predicting protein—protein interactions from primary structure", Bioinformatics 2001, 17:455–60

[2] R. Karchin, K. Karplus, and D. Haussler, "Classifying G-protein coupled receptors with support vector machines". Bioinformatics 2002, 18:147–59.

[3] T. Joachims, Text Categorization with Support Vector Machines: Learning with Many Relevant Features. (Date: 22/01/06)

[4] V. Vapnik, (1998). Statistical Learning Theory. John Wiley & Sons, Inc., New York.

[5] B. Schölkopf, A. J. Smola, "Learning with kernels: support vector machines, regularization, optimization, and beyond", In Adaptive computation and machine learning Cambridge, Mass. MIT Press; 2002:xviii.

Conformational changes induced by the ATP-cycle in ABC transporters: Insights from molecular dynamics simulations

A. Sofia F. Oliveira[1], António M. Baptista[1], and Cláudio M. Soares[1]

[1] *Instituto de Tecnologia Química e Biológica, Universidade Nova de Lisboa, Oeiras, Portugal (asfo@itqb.unl.pt)*

ATP-Binding Cassette (ABC) transporters are ubiquitous membrane proteins that use the energy from ATP-binding or/and hydrolysis to actively transport substrates (named allocrites) across membranes against the concentration gradient [1]. The allocrites are chemically very diverse, ranging from small ions to polypeptides. These transporters are found virtually in all living organisms and mutations in several members of this family have been associated with genetic diseases in humans (such as cystic fibrosis [2] and Tangier disease [3]) and to multidrug resistance in bacteria, fungi, yeasts, parasites and mammals.
ABC transporters can be divided in two groups: importers, which translocate allocrites to the cellular interior, and exporters, which do the opposite. Independently of the transport directionality, ABC transporters are usually composed by a minimum ''functional core'' formed by four modules [1]: two transmembrane domains (TMDs) and two catalytic domains (NBDs).

However, and despite the large amount of experimental and theoretical data available for several family members many fundamental questions about the ABC-family remain unanswered until this moment. In particular, it is still not clear which are the conformational changes induced by ATP-hydrolysis in the NBDs, nor how these rearrangements are ''transmitted'' to the TMDS, in order to drive allocrite translocation.

The main objective of this work is to identify and map the structural changes occurring during an ATP-hydrolytic cycle in three distinct systems: an isolated NBD dimer from *Methanococcus jannaschii* (MJ0796) [4], a full length ABC exporter from *Staphylococcus Aureaus* (Sav1866) [5] and a complete ABC import system from *Escherichia coli* (MalFGK$_2$E) [6]. In this work, we present the results of three computational studies [4-6] using extensive Molecular Dynamics (MD) simulations of several intermediates states of the ATP-cycle. Our MD simulations allowed us to identify that the conformational rearrangements occurring during the ATP-cycle are not restricted to the NBDs, but extend to the TMDs external regions. Additionally, in the context of the complete transporters, we were also able to identify the atomic details associated with the NBD dimer opening upon hydrolysis. Lastly, we suggest a general mechanism for coupling hydrolysis and energy transduction to allocrite translocation (independently of the transporter directionality), in which the NBDs "helical sub-domain region" and the TMDs ''coupling helices'' are the keystones.

[1] C.F. Higgins *Annu Rev Cell Biol* **8** (1992) 67.
[2] J.R. Riordan et al. *Sciencel* **245** (1989) 1066.
[3] M. Dean et al. *Genome Res* **11** (2001) 1156.
[4] A. S. F. Oliveira et al. *J Phys Chem B* **114** (2010) 5486.
[5] A. S. F. Oliveira et al. *Proteins* **79** (2011) 1977.
[6] A. S. F. Oliveira et al. *PLoS Comput Biol* **7** (2011) e1002128.

Interactome3D: a resource for the structural annotation of protein-protein interaction networks

Roberto Mosca[1] **and Patrick Aloy**[1,2]

1. Institute for Research in Biomedicine, Joint IRB-BSC Program in Computational Biology,
c/ Baldiri i Reixac 10-12, 08028 Barcelona, Spain.
2. Institució Catalana de Recerca i Estudis Avançats (ICREA), Spain.

The emergence of high-throughput proteomic initiatives for the identification of new protein interactions and macromolecular complexes has led to the production of large interactome maps for several model organisms. The information contained in these interaction networks is mainly of binary nature, indicating whether one protein interacts with another. However, to fully understand how a complex system like the cell works we need to know the details of how these interactions are performed. 3D high resolution structures of proteins and complexes are one of the best sources of information to extract and interpret such details.

For this reason we created a pipeline for the automatic structural annotation of interactome networks of entire organisms. The pipeline collects all experimentally identified protein-protein binary interactions from several public databases and searches the Protein Data Bank for 3D experimental structures of both single proteins and interactions. In order to improve the structural coverage homology models for single proteins are downloaded from Modbase while models of protein-protein interactions are produced using Modeller.

By applying the pipeline to seven organisms (E. coli, fly, H. pylori, human, mouse, worm and yeast) we observed that the structural coverage of interactomes is very diverse among the different organisms. For a considerable part of the proteins in the interactomes we can find structural data while the structural coverage for the interactions is much lower. Interestingly, for several organisms there is a large number of interactions for which we have structures for both the interactors but not for the interaction itself opening the door to the application of computational prediction methods like protein-protein docking.

Show-EM with PeppeR

Macías JR[12], Poza-Ballesteros A[1], Jiménez-Lozano N[1,2], Carazo JM[1,2]

[1] Spanish National Center for Biotechnology (CNB-CSIC). Darwin, 3. 28049 Madrid, Spain.
[2] GN7 of the National Institute for Bioinformatics (INB).

The Distributed Annotation System (DAS) was designed as a lightweight system for integrating data from a number of heterogeneous, distributed biological databases [1]. The DAS system consists of a set of annotation "layers", each of them containing a particular kind of annotation produced by a specific research laboratory or established database (the annotation provider). A data exchange standard: the DAS XML specification was designed to enable the DAS servers to provide the annotation layers in real time and a DAS client to overlay them producing a single integrated view.

PeppeR [2] is a Java-based, graphical DAS client, designed for visualization of hybrid models (3D-EM volume maps plus atomic resolution structures from fitting experiments) together with all their relevant annotations. PeppeR implements DASx3DEM, an extension of the DAS protocol that allows sharing annotations about hybrid models. Annotations for protein sequences, atomic coordinates, and 3D-EM volumes are collected and displayed through a single graphical visualization interface that provides the users an integrated view of all the annotations together with the hybrid model.

In the current version of PeppeR many features have been updated and improved, especially those regarding the 3D display of the hybrid models, after the release of the AstexViewer library as open source. Also, a new utility tool has been developed: **Show-EM** (*http://biocomp.cnb.csic.es/das/Show/EM/*), allowing PeppeR to be easily integrated in any web portal. Show-EM is a RESTful Web Service that provides a Java Web Start engine for launching PeppeR directly from a Web link. A jnlp file (a small, simple XML file with instructions on how to start Pepper) is dynamically generated and a link to the file is provided. Any Web portal that wants to provide a direct access to PeppeR with a particular hybrid model pre-loaded, only needs to include the provided link in a convenient way.

[1] Dowell et al. BMC Bioinformatics 2001, 2:7, doi:10.1186/1471-2105-2-7.
[2] J R Macías et al. Journal of structural biology 158 (2007) p. 205-13. PMID: 17400476; doi: 10.1016/j.jsb.2007.02.004.

Experimental and Computational Studies Indicate Specific Binding of pVHL Protein to Aurora-A Kinase

Imen Ferchichi[1], Nejla Stambouli[1], Raja Marrackchi[1], Yannick Arlot[2], Claude Prigent[2], Ahmed Fadiel[3], Kunle Odunsi[3],| Amel Ben Ammar Elgaaied[1], and Adel Hamza[1,3,4.]

[1] Laboratory of Genetics, Immunology and Human Pathology, Faculty of Sciences of Tunis, Tunisia,
[2] UMR 6061 Faculty of Medicine of Rennes 1, France,
[3] The Bioinformatics and Computational Biology InitiatΝe, Meharry Medical College, NashVille, Tennessee 36176,
[4] DΝision of Gynecologic Oncology, Department of Surgical Oncology, Roswell Park Cancer Institute, Buffalo, New York

Overexpression of Aurora-A kinase is commonly detected in many cancers [1], whereas the von Hippel-Lindau protein (pVHL) is frequently mutated or absent in renal cell carcinoma [2] and is involved in the Ub proteasome complex, an important degradation pathway. In order to establish a link between Aurora-A overexpression and lack of pVHL protein, we hypothesized that pVHL regulates Aurora-A expression through a physical interaction. We present the first evidence, from both biological assays and computational biology techniques, that human pVHL binds strongly to Aurora-A kinase. Extensive molecular modeling, docking, and dynamic simulations demonstrate that the structure of the pVHL protein would allow it to bind to the TPX2 binding region of Aurora-A [3]. In view of Aurora-A's importance as a therapeutic target for the treatment of cancer, this observation provides novel insights into the Aurora-A/pVHL pathway. In addition, the detailed Aurora-A/ pVHL binding structure obtained will be valuable for the design of future Aurora-A inhibitors as therapeutic agents.

[1] JJ. Zhu et al. *Cancer Genet. Cytogenet.* **159** *(*2005) 10–17.

[2] H. Brauch et al. *Cancer Res.* **60** (2000) 1942–1948.

[3] A. Bird et al. *J. Cell Biol.* **182** *(*2008) 289–300.

Scipion: Towards an electron microscopy integration

A. Quintana[13], J. Vargas[3], A. Poza-Ballesteros[3], J.R. Macías[3], R. Marabini[2], C.O.S. Sorzano[3] and J. M.Carazo[13]

[1]*GN7 of the National Institute of Bioinformatics, Madrid, Spain*
[2]*Computer Science Dept., Univ. Autonoma de Madrid, Madrid, Spain*
[3]*Instruct Image Processing Center (I2PC), BCU, National Center of Biotechnology (CNB-CSIC), Madrid, Spain*

Computational tools for 3D Electron Microscopy (3DEM) are currently in sufficiently mature state as to be integrated into higher-level workflows composed of several simpler steps. This is in fact the path followed by the major packages in the field (Spider, Eman, Imagic, Xmipp, etc.). Additionally, users normally contrast the output of different algorithms as a way to validate their results. Data conversion, handling of different conventions, reorganization of files, etc. is an issue for the unexperienced user who finds a severe barrier to validate her results with different packages.

Several attempts have already been done towards data, packages and workflow integration. EMDataBank (http://www.emdatabank.org) is probably the most successful project on data integration. Although there are some precedents regarding workflow integration, some of them have suffered from a certain lack of functionality and/or impact, as IPLT [1] and SPARX [2]. On the other hand, Appion [3] is a single platform allowing a real integration of different software package. Currently it is probably the best workflow platform in the 3DEM field although a relatively low performance, environment restrictions and the lack of a friendly installation process are some of the weak points of its current implementation.

In this abstract we present Scipion, an image processing platform for 3DEM which aims to integrate several software packages for the elucidation of 3D structures through a workflow approach. The software will allow the execution of reusable, standardized, traceable and reproducible image processing protocols. These protocols will integrate tools from the main 3DEM software packages (bringing full interoperability among them) and are being built upon a consensus of the structural biology community and the expertise of the main European 3DEM software developers. The platform will allow access to powerful computational facilities such as High Performance Computing (HPC), computer cluster and cloud computing. Scipion is a EU initiative directly managed by the I2PC (http://i2pc.cnb.csic.es/).

Scipion is designed as a decentralized and decoupled application mainly developed in Java connected to an ontological database. As Scipion integrates different, the development of the ontology has been a crucial step. This ontology is currently facilitating the interoperability among the packages and eventually it will foster standardization in the 3DEM field. Users can access Scipion from a desktop application (Scipion Desktop Client, SDC), which will provide execution and analysis functionalities, connected to a web services platform. The web services platform can be defined as the Scipion brain as it is in charge of the main tasks: management of the user requests, connection to the workflow engine, data exchanging process between the data storage machine and the execution machine, the DB related chores, etc. Activiti (based on BPMN, a graphical notation to design workflows), acting as the Scipion workflow engine, will invoke sequentially the 3DEM algorithms through some Python wrappers. Python wrappers will be sent to the execution host and will convert the input and output parameters to the required formats and call the specific program.

[1] Philippsen et al. *J. Structural Biology*, **157** (2007) 28-37

[2] Hohn et al. *J. Structural Biology*, **157** (2007) 47-55

[3] Lander et al. *J. Structural Biology*, **166** (2009) 95-102

A MACHINE LEARNING APPROACH TO EXPLORE THE CORRELATION BETWEEN METABOLISM AND ENVIRONMENT IN PROKARYOTES

C. Higuera[1*] , G. Pajares[2], F. Morán[1], J. Tamames[3]

[1] Dpto. de Bioquímica y Biología Molecular I, Facultad de Ciencias Químicas, Universidad Complutense Madrid, Avd. Complutense S/N, 28040, Madrid, Spain

[2] Dpto. de Ingeniería de Software e Inteligencia Artificial, Facultad de Informática, Universidad Complutense de Madrid, 28040 Madrid, Spain

[3] Centro Nacional de Biotecnología, CSIC, 28049, Madrid, Spain

[*]Email:clarah@solea.quim.ucm.es

It has been proposed that microbial species and taxa have environmental preferences, that is, they favor particular habitats amongst others. The understanding of the connection between the functional traits conferred by their genome complement and the environment is critical for understanding the structure and functioning of natural microbial communities. To this end, we have explored the relation between metabolism and environmental preferences, to address two main points: first, the existence of environmental-associated genes, which could be responsible of conferring a better adaptation to the particular physicochemical conditions shaping each environment. Second, the possibility of building learning classifiers that could use metabolic information, presence or absence of genes and pathways, to predict the environmental preferences of bacteria.

Machine learning has proven to be very useful in the field of bioinformatics [1] allowing the extraction of underlying information in data. In this work we present the results of a machine learning approach based on SOM (Self Organizing Map) used to achieve the aforementioned goals. Preliminary tests with this method show that it is possible to classify species by their metabolism into topologically ordered groups on a map and visualize which ones share common environments. Using this method we will be able to predict the environments in which an unknown bacteria may live. This technique combined with other machine learning methods such as fuzzy clustering or learning vector quantization could provide us with valuable information for understanding the functional basis of the adaptation of bacteria to different habitats.

Acknowledgements

This work has been supported by Grant BFU2009-12895-C02-02, MICINN, Spain

[1] P. Larrañaga et al. Machine learning in bioinformatics, *Briefings in Bioinformatics*, vol. 7, p.86 , 2006.

miRGate : A web based tool to extract relevant relationships amongst user-defined microRNAs profiles and gene expression.

Eduardo Andrés León[1], Daniel Glez. Peña[2], Gonzalo Gómez[1] and David G. Pisano[1].

[1]Bioinformatics Unit (UBio), Structural Biology and Biocomputing Programme, Spanish National Cancer Research Centre (CNIO), Madrid, Spain
[2]Higher Technical School of Computer Engineering, University of Vigo, Ourense.

MicroRNAs (miRNAs) have been found to play an important role as regulatory molecules in several biological processes, they are believed to be implicated in cancer development and regulating several processes such as cell cycle control, metastases development. In addition, impaired miRNA-target regulation have been linked to neurological disorders and cardiovascular disease.

Nowadays, current techniques to determine and characterize the role of miRNAs in the molecular basis of human disease malignancies include high-throughput approaches, such as microarrays or next-generation sequencing techniques. These approaches provide miRNAs signatures and large catalogues of candidate targets whose direct biological relationships have to be elucidated. MiRGate takes these catalogues as input to achieved the most probably miRNA-Transcript target pair based on predictions made by 5 different prediction methods: miRanda[1], Pita [2], RNAHybrid[3], Microtar[4] ,targetscan[5]. As targetscan place importance on the conservation, EnsEMBL alignments have been used to find out if a target is conserved between species.
Experimentally validated targets, although are expensive and time consuming are needed to understand the implication of predicted targets and provide valuable information to distinguish weak predictions. To increase our understanding in miRNA-UTR targets, three different databases with experimentally confirmed targets were selected. They are frequently updated and manually curated from papers. Tarbase[6], miRTarbase[7] and miRecords[8].

Our approach represents an extension of traditional overrepresentation methods employed for functional analysis of gene lists. The procedure consists of the application of recursive Fisher's exact tests for 2×2 contingency tables containing the number of microRNA-targets associations found in both lists of interest compared to a set of reference (e.g. the rest of genes in the array). Thus, a p-value is assigned to each microRNA together with False Discovery Rate correction (FDR) to account for multiple testing. The output of the analysis consists on a subset miRNAs showing FDR < 0.05 after Fisher's exact test. These microRNAs are selected on the basis of their non-random association with the gene expression list of interest.

MiRGate has successfully been used in some projects, and it was a very helpful tool for finding Epstein-Barr virus microRNAs that repressed BCL6 expression in diffuse large B-cell lymphoma [9].

[1] Enright AJ et al. *Genome Biology* **5** (2003) R1.
[2] Michael Kertesz et al. *Nature Genetics* **39** (2007) 1278–1284.
[3] Marc Rehmsmeier et al. *RNA* **10** 1507-1517
[4] Rahul Thadani et al. *BMC Bioinformatics* **5** (2006) Sup5:S20.
[5] Benjamin P Lewis et all. *Cell* **120** (2005) 15-20
[6] Praveen Sethupathy et al. *RNA* **12** (2006) 192-197
[7] Hsu SD et al. *Nucleic Acids research* **39** (2010) 163-169.
[8] Xiao F et al. *Nucleic Acids research* **37** (2008) 105-110.
[9] D Martín-Pérez et al. *Leukimia* (2011).

Disease-specific network-based method for candidate genes prioritization

Luz Garcia-Alonso[1]*, Verónica Llorens[1]*, Enrique Vidal[1], Jose Carbonell[1], Ignacio Medina[1] and Joaquin Dopazo[1].

[1] Bioinformatics and Genomics Department, Centro de Investigación Príncipe Felipe (CIPF), Valencia, Spain.
** These authors contributed equally to this work.*

The identification of disease-associated genes has long been a necessary labour toward enhancing our knowledge of disease mechanisms, diagnosis, prognosis and therapy. The development of genome-scale techniques such as SNPs and CNV genotyping or, more recently, exome sequencing offers a step forward in this purpose. However, these techniques generate a large amount of disease-candidate genes, making experimental corroboration laborious and expensive. Under the assumption that physical and functionally related genes give rise to identical or related phenotypes [1] but the genes relationship differs across diseases [2], we introduce a disease-specific network-based method for candidate genes prioritization. Prioritization methods previously proposed assume that disease-associated genes have to be related in an arbitrary way, i.e. to be close in the protein-protein interaction network [3]. In contrast, our method search for specific traits defined by the disease-associated genes in a wide range of physical and functional networks and network measurements. These disease trademarks establish the rules to prioritize the disease-candidate genes. The method was validated using the leave-one-out cross-validation (LOOCV) strategy. For this purpose, 6470 gene-diseases relationships from Online Mendelian Inheritance in Man (OMIM), ORPHANET, Uniprot and Ensemble databases were collected and related through the Disease Ontology (DO) and the Human Phenotype Ontology (HPO). Our results provide evidence that disease-associated genes profiles vary across diseases, which prove the need for a disease-specific method. Moreover, the ability of establishing disease relationships though the DO and the HPO allows our method to deal with diseases poorly characterised at genetic level.

[1] I. Felfman et al. *Proceedings of the National Academy of Sciences* **105** (2008) 4323-4328.
[2] S. Köhler et al. *The American Journal of Human Genetics* **82** (2008) 949-958.
[3] X. Wang et al. *Briefings in Functional Genomics* **10** (2011) 280-293.

Expression quantitative trait loci mapping from multivariate expression profiles

Inma Tur[1,2], and Robert Castelo[1,2]

[1]Dept. Experimental and Health Sciences, Universitat Pompeu Fabra, Barcelona, Spain.
[2]Research Program on Biomedical Informatics, Institut de Recerca Hospital del Mar, Barcelona, Spain.

The simultaneous assay of genetic variation and gene expression allows one to map in the genome the so-called expression Quantitative Trait Loci (eQTL), which are genetic loci harboring DNA variants that affect quantitative levels of gene expression from specific genes. Differently to classical QTL mapping of a single phenotype, such as blood pressure, gene expression data is a multivariate vector involving thousands of molecular phenotypes, one of each probed gene. The finely tuned control exerted on the gene expression pathway by regulatory mechanisms renders expression levels of functionally related genes, highly correlated. This poses a major challenge to classical QTL mapping univariate techniques, which is to distinguish direct from indirect associations between gene expression profiles and genetic loci.

We introduce a new methodology for eQTL mapping that employs a multivariate model of gene expression whose underlying model, based on conditional Gaussian distributions and mixed graphical model theory [1], explicitly deals with direct and indirect associations. The method relies on an exact test for the conditional association between a genetic marker and an expression profile, which according to our simulated data, plays a crucial role to appropriately control false positive predictions in such a multivariate context. Additionally, the methodology can control for confounding effects in expression data by using a fixed effects model based on surrogate variable analysis [2].

We have assessed the performance of the proposed method in a widely tested yeast data set [3] where two different strains are crossed to generate 112 segregants, which were profiled in their gene expression (6,216 genes) and genotyped (2,906 markers). We show that it predicts the strongest associations between genetic loci and expression profiles from genes whose function (according to Gene Ontology) is related to the genes in *cis* to the genetic loci in a higher degree than the genes predicted by a classical QTL mapping method such as single marker regression. In the particular case of the eQTL hotspot located chr3:90790, we observe that while the top-three genes predicted by single marker regression have no straightforward functional relationship to the genes within the 1kb region of the eQTL, our method assigns the strongest association to LEU2, a *cis* gene immediately downstream of the eQTL which is then located in its promoter region and which is involved in the leucine biosynthesis pathway. Likewise, it assigns second and third stronger associations to LEU1 and OAC, which are genes with binding sites of LEU3 in their promoter region, a major regulator switch of this pathway [4].

[1] Lauritzen, S. and Wermuth, N. *Ann Stat* **17** (1989) 31-57.
[2] Leek, J.T. and Storey, J.D. *PLoS Genet* **3** (2007) 1724-35.
[3] Brem, R.B. and Kruglyak, L. *Proc Natl Acad Sci USA* **102** (2005) 1572-7.
[4] Chin, C.S. *et al. PLoS Biol* **6** (2008) e146.

SG-VarHunter: An efficient strategy for variant calling and annotation applied to color-space data

Triviño JC[1], Rosa-Rosa JM[1], Rodríguez-Cruz O[1], Cabo-Díez M[1], Collado C[2], Fernández-Pedrosa V[2], Pérez-Cabornero L[3], Santillán S[3], Zúñiga S[1*]

[1,2,3] Departments of Bioinformatics[1], New Technologies[2] and Medical Genetics[3]. Sistemas Genómicos S.L., Valencia, Spain

Advances in Next-Generation Sequencing chemistries and sample preparation protocols are rapidly transforming today's Biology. The new generation of sequencers produces large amounts of data, for one or a set of samples in parallel, in a very short time and at a reasonable cost per base [1]. One of the most successful applications of NGS into Genetic Diagnostics has been the combination of Next-Generation Sequencing and targeted enrichment strategies [2-3], which allows the efficient characterization of a set of genes or even the entire exome. This approach has been shown to be extremely valuable in the identification of novel genes responsible for rare mendelian disorders [4-6].

A critical issue in the successful application of NGS into Genetic Diagnostics is the ability to correctly identify variants from such massive datasets never achieved by any previous technology. The rapid evolution of NGS platforms has force the development of new analysis algorithms, however the combination of different pieces of software provides very different results and so the establishment of an efficient and accurate pipeline to call variants is very challenging.

Here we describe a robust and validated strategy for variant identification that we have called "SG-VarHunter". This work also includes a reliable comparison between the two commercially available color-space sequencer versions, SOLiD4 and SOLiD5500XL and shows how improvements in the chemistry have minimized the false negative variant rates.
The method here presented has shown to be precise and reliable for variant calling and is currently used in our laboratories for Genetic Diagnostics in genetic heterogeneous diseases.

[1] Mardis ER. *Nature.* **470(7333)** (2011) 198-203.
[2] Mamanova L et al. *Nat Methods.* **7(2)** (2010) 111-118.
[3] Koboldt DC et al. *Brief Bioinform.* **11(5)** (2010) 484-498.
[4] Bamshad MJ et al. *Nat Rev Genet* **12(11)** (2011) 745-755.
[5] Majewski J et al. *J Med Genet.* **48(9)** (2011) 580-589.
[6] Ku CS et al. *Hum Genet.* **129(4)** (2011) 351-370.

Keywords: NGS, Exome, Targeted resequencing, Genetic Diagnostics, Variant calling, Color-space.

Adding supplementary data for a better understanding of gene expression in kernel reduced dimension representations

E. Vegas[1], F. Reverter[1], J.M. Oller[1]

[1] *Department of Statistics. University of Barcelona. Diagonal, 643. 08028 Barcelona.*

Since microarray data show generally nonlinear behavior, finding methods that can handle such data is of great importance if as much information as possible is to be gleaned. Kernel representation offers an alternative to nonlinear functions by projecting the data into a high-dimensional feature space, which increases the computational power of linear learning machines [1].

Kernel methods enable us to construct different nonlinear versions of any algorithm which can be expressed solely in terms of dot products. Thus, kernel algorithms avoid the explicit usage of the input variables in the statistical learning task. Kernel machines can be used to implement several learning algorithms but they usually act as a black-box with respect to the input variables. This could be a drawback in biplot displays in which we pursue the simultaneous representation of samples and input variables.

In this work we develop a procedure for enrich the interpretability of Kernel PCA by adding in the plot the representation of input variables. In particular, for each input variable (gene) we can represent locally the direction of maximum variation of the gene expression. Our implementation enables us to extract the nonlinear features without discarding the simultaneous display of input variables (genes) and samples (microarrays).

[1] Scholkopf, B.; Smola, A.J. (2002). Learning with Kernels - Support Vector Machines, Regularization, Optimization and Beyond. Cambridge,MA. MIT Press.

The Human Transcriptome: Genomic Sites of Origin, Processing Fates and Subcellular Localization

The ENCODE consortium and Sarah Djebali[1]

[1] Bioinformatics and Genomics, Centre for Genomic Regulation (CRG) and UPF, Barcelona, Catalonia, Spain

The biogenesis of the RNA products found in all cells involves polymerase-directed transcription, transcript processing, modification, transport and ultimately degradation. This study develops comprehensive genome-wide catalogues of multiple types of transcribed elements, such as transcription start sites (TSS), exons, splice sites, transcripts and genes. These elements are derived from long (>200 nucleotides [nts]) polyadenlyated (poly A+) and non-polyadenylated (poly A-) transcripts identified in the nuclear, sub-nuclear (chromatin, nucleolus, nucleoplasm) and cytosolic sub-compartments from 15 primary and cancer cell lines. Analyses of these catalogues confirm pervasive transcription across the genome resulting in significant increases in transcription start site, exon, splice site, transcript and gene elements and a reduction in the median length of intergenic regions. Genomically encoded transcripts before and after processing have been monitored resulting in a steady-state genealogy for many RNAs that can be followed into cellular sub-compartments. These data also provide insights into 1) the extent and sub-cellular compartmentalization of many annotated and novel RNAs, 2) cell type and compartment specific transcript isoform prevalences and TSSs, 3) the extent, characterization and localization of RNA editing and 4) the identification of specific class of transcriptional enhancers distinguished by the RNAs emanating from these genomic regions.

A web tool for annotation of draft conifer genomes using Maker

Pedro Seoane[1], Rocío Bautista[2], Noé Fernández[1,], Darío Guerrero-Fernández[2] y M. Gonzalo Claros[1,2].

[1]Departamento de Biología Molecular y Bioquímica, Facultad de Ciencias, Universidad de Málaga, Málaga
[2]Plataforma Andaluza de Bioinformática-Centro de Supercomputación y Bioinformática, Universidad de Málaga, Malaga

Pinus pinaster is an economically and ecologically important species that is becoming a woody gymnosperm model. Its enormous genome size makes whole-genome sequencing approaches hard to apply. *P. pinaster* is a conifer, which are separated from angiosperms by more than 300 million years of independent evolution so the current reference models are not as valid as in angiosperms. In fact, studies on the conifer genome are revealing unique information which cannot be inferred from currently sequenced angiosperm genomes (such as poplar, *Eucaliptu*s, *Arabidopsis* or *Oryza*): around 30% of conifer genes have little or no sequence similarity to plant genes of known function. And last, the pine genome is replete with highly repetitive, non-coding sequences. Therefore, a dedicated annotation tool seems to be required for conifers.

We have developed an easily configurable genome annotation tool called GeNOTE [1] which has been used to accurately annotate several BAC clones from a *P. pinaster* library [2]. Genomes are being sequenced at a far rate than they are being annotated and GeNOTE has not been designed to annotation of whole genomes. Other tool described to annotation of small whole genomes is called MAKER [4] that identifies repeats, aligns ESTs and proteins to a genomic sequence and produces 'in-silico' gene predictions. This tool has been tested in several BACs of *Pinus taeda* using monocot and dicot plant sequences as training models. From the above description, we think that pine gene prediction must be done with conifer training datasets to avoid false-annotations and do not skip putative exons. We have set up MAKER with pine ESTs from EuroPineDB [3] and Pine Gene Index v 8.0, and conifer proteins from Swiss-Prot and TrEMBL, optimizing the process of annotation for *P. pinaster.* in the seek of accurate annotations. Also, we have developed some scripts to simplify the output file of MAKER and show some statistics as descriptions related to gene, complete/uncompelte gene, gene lengths, etc. On the other hand, the installation and configuration of MAKER is not trivial, so we have developed a web server to MAKER. This web tool is allowing the accurate annotation of great fragments of pine genome (http://www.scbi.uma.es/maker, soon available).

References:

1.-Fernández-Pozo N., Guerrero-Fernández D., Bautista R., Gómez-Maldonado J., Avila C., Cánovas FM., Claros MG. **GeNOTE: a web tool for annotation of non-model, eukaryotic, unfinished sequences**. ISBN: 978-84-614-4481-6 Pg. 74, 2010

2.- Rocío Bautista, David P. Villalobos, Sara Díaz-Moreno, Francisco R. Cantón, Francisco M. Cánovas, M. Gonzalo Claros. **New strategy for pinus pinaster genomic library construction in bacterial artificial chromosomes**. Investigación agraria. Sistemas y recursos forestales **17**, 238–249. 2008.

3.- Fernández-Pozo N., Canales J., Guerrero-Fernández D., Villalobos D.P., Díaz-Moreno S., Bautista R., Flores-Monterroso A., Guevara M.A., Perdiguero P., Collada C., Cervera M.T., Soto A., Ordás R., Cantón F.R., Avila C., Cánovas F.M. and Claros M.G. **EuroPineDB: a high-coverage Web database for maritime pine transcriptome.** BMC Genomics 2011, **12**:366 . 2011.

4.-Cantarel BL, Korf I, Robb SM, Parra G, Ross E, Moore B, Holt C, Sánchez Alvarado A, Yandell M. **MAKER: an easy-to-use annotation pipeline designed for emerging model organism genomes**.
Genome Res. 2008 Jan;**18**(1):188-96,2007.

Intron/exon boundary prediction for SNP discovery from transcriptome assemblies

Darrell Conklin[1], Aitor Albaina[2], Iratxe Montes[2], Pablo Markaide[2], Iratxe Zarraonaindia[2], Andone Estonba[2]

[1]Department of Computer Science and Artificial Intelligence, University of the Basque Country, and IKERBASQUE, Basque Foundation for Science
[2]Department of Genetics, Physical Anthropology and Animal Physiology, University of the Basque Country

Overview. Next generation sequencing (NGS) technologies have led to a revolution in population genetics, notably in the discovery of DNA sequence variation through high-throughput discovery of single nucleotide polymorphisms (SNPs) [1]. In a typical NGS sequencing project thousands of SNPs will be called even with conservative parameters, and given the cost and effort involved in genotyping (for all selected SNPs, experimentally determining the alleles occurring within a population), high SNP success rates (verified polymorphic sites) and low genotyping failure rates are essential. A central reason for genotyping failures when SNPs are called from transcriptome assemblies is the unanticipated proximity of intron/exon boundaries (IEBs) in the areas of genotyping primers [2,3]. To avoid IEBs, one can either validate only a small set of SNPs within those genes that are already known [4], or attempt to map contigs to a close annotated genome if one exists [5]. A substantially more complex scenario arises when contigs cannot be reliably mapped to known genomic sequences at the nucleotide level [2,6].

Sequencing and bioinformatics. European anchovy transcripts from a pool of ten individuals and four tissues were sequenced using Roche 454 GS-FLX technology, assembled into contigs using Roche gsAssembler, and high confidence SNPs were called using the Roche gsMapper tool. The Ensembl Perl API (v64) was used to derive a database of coding sequences for all transcripts from the Ensembl zebrafish genome (Zv9), each further annotated with all known IEBs in protein coordinates. Contigs from assembled anchovy reads were mapped at the protein level using blastx with additional logic applied to identify a possible unique orthologous zebrafish gene. For each SNP in contigs with an identified ortholgous gene, and for each HSP with all transcripts of that gene, corresponding SNP positions in the zebrafish transcript could be calculated based on HSP positions. These results are filtered to retain only those SNPs that are not in proximity to any IEB or proximal to untranslated regions where IEBs may be less conserved between orthologous genes.

Discussion. The method identifies those coding SNPs which are not close to a predicted IEB, as inferred from all transcripts from a potentially orthologous gene. The solution developed here differs from a recent approach [3] in the detection of orthologues at the more sensitive protein sequence level, and in the consideration of IEBs in all splice variants using the Ensembl gene/transcript hierarchy. The predictions of the method will soon be validated against the anchovy genome which is currently being sequenced. The method reported here should be applicable in the future to SNP discovery in other non-model species organisms.

[1] Garvin, M. et al. (2010) Molecular Ecology Resources, **10**:915.
[2] Wang, S. et al. (2008) BMC Genomics, **9**:450.
[3] Milano, I. et al. (2011) PLoS ONE, **6**(11).
[4] Meyer, E. et al. (2009) BMC Genomics, **10**:219.
[5] Hoffman, J. (2011) Molecular Ecology, **11**:703.
[6] Iorizzo, M. et al. (2011) BMC Genomics, **12**:389.

Clustering of DNA words and biological function

José L. Oliver, Michael Hackenberg, Guillermo Barturen, Francisco Dios, and Antonio Rueda

Dpto. de Genética, Facultad de Ciencias, Universidad de Granada, Campus de Fuentenueva s/n, 18071-Granada, Spain & Lab. de Bioinformática, Centro de Investigación Biomédica, Campus de la Salud, Avda. del Conocimiento s/n, 18100-Granada, Spain

Relevant words in literary texts (key words) are known to be clustered, while common words are randomly distributed [1,2]. Our group [3] has proven that this principle is successful in detecting semantic meaning (keyword extraction) in conventional literary texts, as well as in comma-less texts (i.e. texts from which all spaces and punctuation marks have been removed). Given the clustered distribution of many functional genome elements [4,5,6,7,8,9,10,11], we hypothesize that the biological text per excellence, the DNA sequence, might behave in the same way: *k*-length words (*k*-mers) with a clear function may be spatially clustered along the one-dimensional chromosome sequence, while less-important, non-functional words may be randomly distributed.

To explore this linguistic analogy, we calculate a clustering coefficient for each *k*-mer (k = 2-9 bp) in human and mouse chromosome sequences, then proving if clustered words are enriched in the functional part of the genome [12]. First, we show that word clustering is positively related to the enrichment within exons and TFBSs, while a much weaker relation exists for repeats, and no relation at all exists for introns. Second, enrichment/depletion experiments show that highly clustered words are significantly enriched in exons and conserved TFBSs, while they are depleted in introns and repetitive DNA. Noteworthy, the percentages of enrichment for unclustered words are always lower than for clustered words. Lastly, we developed an algorithm to extract vocabularies of true 'DNA keywords', which include words of different lengths, and are free of inclusions and redundancies.

The clustering of DNA words may be a novel principle to detect functionality in non-annotated chromosome sequences. As evolutionary conservation is not a prerequisite, the proof of concept described here may be the first step towards new ways to detect species-specific functional DNA sequences and the improvement of gene and promoter predictions, thus significantly contributing to the current quest for function in the genome.

[1] Ortuño M *et al. Europhysics Letters* **57** (2002) 759-764.
[2] Zhou H *et al. Physica A* **329** (2003) 309-327.
[3] Carpena *et al. Phys. Rev E.* 79 (2009): 035102(R) (1-4)
[4] Kendal WS. *BMC Evol Biol* **4** (2004) 3.
[5] Neel JV. *Blood* **18** (1961) 769-777.
[6] Durand D *et al. Comput Biol* **10** (2003) 453-482.
[7] Bird AP. *Nature* **321** (1986) 209-213.
[8] Hackenberg M *et al. BMC Bioinformatics* **7** (2006) 446.
[9] Boeva V *et al. Algorithms Mol Biol* **2** (2007) 13.
[10] Berman BP *et al. Proc Natl Acad Sci U S A* **99** (2002) 757-762.
[11] Sargsyan K *et al. Nucleic Acids Res* **38** (2010) 3512-3522.
[12] Hackenberg *et al. J Theor Biol* (submitted).

Portray of the transcriptome of human mesenchymal stem cells using RNA-seq

Beatriz Rosón[1], Consuelo del Cañizo[2], Fermín Sánchez-Guijo[2] and Javier De Las Rivas[1].

[1]Bioinformatics and Functional Genomic Group, Cancer Research Center (CiC-IBMCC, CSIC/USAL), E37007 Salamanca. Spain. [2]Haematology Department. Hospital Universitario de Salamanca. Spain.

By definition, Stem Cells (SC) own the potential for self-renew and differentiate into other cell types, giving rise to more specialized tissues. Recently discovered, Mesenchymal Stromal/Stem Cells (MSC), represent a powerful tool in human cell therapy, capable to differentiate into bone, cartilage, muscle, or fatty cells (1). Human MSC can be found immerse into adipose tissue, placental tissue, or into the highly complex and specialized bone marrow microenvironment, where they coexist and support Haematopoietic Stem Cells (HSC). Their gene expression levels are partly consequence of the surrounding atmosphere, but also responsible for their cellular behaviour and decision-making, and a field still in need to be deciphered and well defined. The intensively studied hematopoietic niche together with the rest of tissues acting as home for the MSC, provide a wide range of functional states to approach a comparative study of their transcriptome.

We have used RNA sequencing technologies, to undertake the study of the human MSC transcriptome. We would present here how different groups of genes are distributed along with their abundance values, and determine which genes are switched on/off in the cells. Some functional groups of genes present common expression profiles in all MSC analyzed, and others appear differentially expressed, picturing an immature gene signature or, by contrast, a tissue specific pattern that belongs to a functionally committed hMSC.

In order to mine the big amount of data coming from RNA-seq experiments, we have applied different available algorithms on the assembly and counting steps, (such as *Cufflinks* (2), *HTSeq-counts* (3), or *DESeq* (4)). We worked also under the R/Bioconductor framework to obtain a robust statistical analysis and graphical support.

References

1. Friedenstein, A.J., Piatetzky-Shapiro, I.I., Petrakova, K.V.: Osteogenesis in transplants of bone marrow cells. *J.Embryol.Exp.Morphol*. 16(3):381-90. (1966)

2. Trapnell C, Williams BA, Pertea G, Mortazavi AM, Kwan G, van Baren MJ, Salzberg SL, Wold B, Pachter L.: Transcript assembly and quantification by RNA-Seq reveals unannotated transcripts and isoform switching during cell differentiation *Nature Biotechnology*. 28, 511–515. (2010)

3. Anders, Simon. HTSeq: Analysing high-throughput sequencing data with Python. (2011) Version 0.5.3. http://www-huber.embl.de/users/anders/HTSeq/doc/count.html#

4. Anders S., Huber W.: Differential expression analysis for sequence counts data. *Genome Biology*. 11:R106. (2010).

Bioinformatic characterization and transcriptomic analysis of synthetases and related biological processes of a newly sequenced bacterial genome, in order to guide the metabolic engineering

Rodríguez-García A[1], Kosalková K[1], Albillos SM[1], Martínez-Castro M[1], Barreiro C[1], Sedano J[2], and Prieto C[1]

[1] Instituto de Biotecnología de León (INBIOTEC), Parque Científico de León, Av. Real, 1, 24006 León, Spain.
[2] Instituto Tecnológico de Castilla y León (ITCL), Pol. Ind. Villalonquéjar, C. López Bravo 70, 09001 Burgos, Spain.

Natural products produced by bacteria, fungi, and plants provide a rich source of compounds with pharmaceutical utilities such as antibiotic, anticancer, immunosuppressant, antiparasitic, antiviral activities. Bioinformatics studies in this area have been restricted for years due to the scarcity of available biological data. However, the great improvement of sequencing technology has enabled the rapid and cheaper sequencing of bacterial genomes. This fact produces a new scenario in microbiology research, in which we have more information than knowledge about the organism. Therefore, new bioinformatics efforts are required to characterize the metabolic clusters and analyze the *omic* data that is being generated.

In this context, the present work has been developed, in which the genome of *S. tsukubaensis* has been sequenced. The first step to gain knowledge about the metabolic pathways is in the identification and characterization of gene clusters which participate in biosynthetic processes. For this purpose, we have implemented a new Bioinformatic tool named NRPSsp [1], which predicts the binding substrate of a given non-ribosomal synthetase sequence. This tool is freely available online at: www.nrpssp.com.

The second step of the project consists of obtaining and analyzing transcriptomic data in order to infer knowledge about the overall operation of the bacteria. Personalized microarrays have been designed to measure the expression levels of genes, sRNAs and antisense transcription regions. These entities have been defined with homology and with the execution of prediction tools. The samples were obtained as temporal series, mapping the most important time points in the production of metabolites. In addition, non-production conditions were also used to obtain samples. Coexpression and differential expression analysis protocols were developed [2] and a new methodology which integrates production values in the analysis was implemented. In this way, the analyses have identified and defined the transcriptional behavior of main sinthetases and processes related with the metabolic activity.

Moreover, the results have opened the door to perform directed genetic modifications and culture medium optimizations in order to improve the production of natural products. All the proposed methodology constitutes an innovative approach adapted to the new landscape in metabolic engineering research [3].

[1] Prieto C et al. *Bioinformatics* (2011) doi: 10.1093/bioinformatics/btr659.
[2] Prieto C et al. *Plos One* **3** (2008) e3911.
[3] Pickens et al. *Annu. Rev. Chem. Biomol. Eng.* **2** (2011) 211-36.

APPRIS, annotating the human genome with protein features

Jose Manuel Rodriguez[1], Paolo Maietta[2], Iakes Ezkurdia[2], Gonzalo Lopez[2,3], Alessandro Pietrelli[1,4], Jan-Jaap Wesselink[2,5], Alfonso Valencia[1,2], and Michael L. Tress[2]

[1]Spanish National Bioinformatics Institute (INB), Spanish National Cancer Research Centre (CNIO), Madrid, Spain.
[2]Structural Biology and Biocomputing Programme, Spanish National Cancer Research Centre (CNIO), Madrid, Spain.
[3]Present address: Center for Computational Biology and Bioinformatics, Columbia University, New York, NY 10032, USA.
[4]Present address: Institute for Biomedical Technologies ITB – CNR, Milan, Italy.
[5]Present address: Molecular Modelling Group, Biomol Informatics, Centro de Biología Molecular "Severo 0choa", Madrid, Spain.

The role played by alternative splicing in the modulation of cellular function is not yet clarified, but recent works have suggested that most alternative isoforms are likely to have altered structure, localisations and cellular functions [1,2]. Given the likely ubiquity in the cell of alternative isoforms, the role of these alternatively spliced gene products is becoming an increasingly important question.

APPRIS is a database that deploys a range of computational methods to provide value to the annotations of the human genome. For example, variants are mapped to known structures using the module MATADOR and *firestar* [3] predicts individual functionally important residues. The database is currently being used to provide genome wide protein feature annotation for the human genome in collaboration with the high quality manual annotations from GENCODE consortium.

As part of the annotation process, the database selects one CDS for each gene as the principal functional variant. The principal isoform for each gene is selected based on the annotations from each of the eight modules implemented in APPRIS. Determining the constitutive variant is a critical first step in the study of the implications of alternative splicing.

We have annotated the 20,700 genes in the current stable Gencode release. Based on this information we have been able to select a principal isoform for 80% of the genes in the human genome. Many of the alternative variants are likely to have changed structure, localisation or function. Over 47% of the annotated alternative splice variants have damaged Pfam functional domains, while over 40% of the alternative variants are likely to have substantially altered protein structure.

Web site: http://appris.bioinfo.cnio.es/.

[1] M. Tress et al. *Proc Natl Acad Sci USA* ***104*** (2007) 5495-5500.
[2] D. Talavera et al. *PLoS Comput Biol.* ***3*** (2007) 33.
[3] G. Lopez et al. *NAR* ***39*** (2011) W235-41.

CorkOakDB: Assembly and annotation of heterogeneous ESTs in a non-model organism

Paulo Almeida[1], Isabel Queirós Neves[1], Aleix Badia[2], Andreas Bohn[2], and José B. Pereira-Leal[1]

[1]Instituto Gulbenkian de Ciência, Oeiras, Portugal
[2]Instituto de Tecnologia Química e Biológica, Oeiras, Portugal

The characterization of the cork oak genome is potentially interesting for economic reasons. This project aims to provide the community with a genomic resource, based on forty five cork oak EST libraries, extracted from different locations, tissues and conditions.

In order to accomplish robustness and flexibility, an EST assembly and annotation pipeline, tightly connected to a database, was developed and optimized through an iterative design process. The pipeline consists of four steps: preprocessing, assembly, protein prediction and functional annotation. Preprocessing involves clustering similar sequences [1], trimming reads (removal of low quality, artifact and poly A/T sequences) [2], removing contaminants (by comparing hits against contaminant and plant databases), masking repeats [3] and extracting chloroplast and mitochondria sequences. The assembly step, using either individual libraries or an assembly thereof (multilibrary assembly) was done with the MIRA software [4], using an optimized set of parameters. Proteins were predicted by the prot4EST program [5] and an Interpro search [6] produced functional annotation for the putative peptides. The results of each pipeline run are versioned for easy retrieval, and intermediate data are kept in the database and file system.

Here we present the final structure of the pipeline, as well as key results that prompted design decisions. These include removal of several pipeline steps and assessment of their influence in the number and length of assembled contigs; evaluation and optimization of the contaminant database; comparison of individual and multilibrary assemblies; rationale for versioning sequences in subsequent pipeline runs.

The knowledge assembled in the database may be explored through a web application, whose main interface is the gene view, featuring sequence data, GO annotations [7], Interpro assignments, KEGG pathways and best BLAST hits [8] against general and plant-specific databases. Genes of interest can be discovered by searching specific fields or by running a nucleotide or protein BLAST search against the Cork Oak database.

[1] B. Niu et al. *BMC Bioinformatics* **11** (2010) 210.
[2] J. Falgueras et al. *BMC Bioinformatics* **11** (2010) 38.
[3] A.F.A Smit et al.http://www.repeatmasker.org (1996-2010).
[4] B. Chevreux et al. *Computer Science and Biology: Proceedings of the German Conference on Bioinformatics* **99** (1999) 45.
[5] J.D. Wasmuth et al. *BMC Bioinformatics* **5** (2004) 187.
[6] S. Hunter et al. *Nucleic Acid Res* **37** (2009) 211.
[7] A. Conesa et al. *Bioinformatics* **21** (2005) 3674.
[8] S.F. Altschul et al. *J Mol Biol* **215-3** (1990) 403.

On the Molecular Evolution of Segmentally Duplicated Sequences

Diego A. Hartasánchez[1], Oriol Vallès-Codina[1], and Arcadi Navarro[1,2,3]

[1]*Institute of Evolutionary Biology (UPF-CSIC), PRBB, Doctor Aiguader 88, 08003, Barcelona, Catalonia, Spain.*
[2]*National Institute for Bioinformatics, Universitat Pompeu Fabra, Barcelona, Spain.*
[3]*Institució Catalana de Recerca i Estudis Avançats (ICREA), Barcelona, Catalonia, Spain.*

Structural Variation (Segmental Duplication and Copy-Number Variation) is a widely observed phenomenon in the genomes of humans and other primates [1], as well as in many other eukaryotic genomes. In humans, CNVs are thought to be associated with disease susceptibility and SDs have been identified as sources of evolutionary innovation [2]. Despite its abundance and relevance, a theoretical description of the underlying forces shaping the evolution of structural variation is still lacking. The proper characterization of the interplay between mutation and gene conversion in duplicated regions is thus fundamental [3]. We have therefore developed a forward-time simulation program that incorporates duplications and focuses on the effect of concerted evolution (the non-independent evolution of duplicated regions) on standard statistical tests that detect regions subject to selection in the genome. We have observed that under realistic biological scenarios, concerted evolution might mimic the effect of both weak purifying selection or weak positive selection, leading to the preliminary conclusion that some CNV regions might have been erroneously identified as singe-copy regions subject to selective forces.

[1] Gazave, E. et al. *Genome Research* **21** (2011) 1626-39.
[2] Marques-Bonet, T. et al. *Trends in Genetics* **25** (2009) 443-54.
[3] Mansai, S.P. et al. *Genes* **2** (2011) 313-31.

Comparative proteomics in Human Genome

Iakes Ezkurdia[1], Angela Del Pozo[1], Adam Frankish[2], Jose Manuel Rodriguez[3], Jennifer Harrow[2], Keith Ashman[1], Alfonso Valencia[1], Michael Tress[1]

[1]Structural Biology and Biocomputing Programme, Spanish National Cancer Research Centre (CNIO), Madrid, Spain
[2] Wellcome Trust Sanger Institute, United Kingdom
[3]Spanish National Bioinformatics Institute (INB), Spanish National Cancer Research Centre (CNIO), Madrid, Spain

Advances in high-throughput mass spectrometry are making proteomics an increasingly important tool in genome annotation projects. We have use public mass spectrometry repositories to provide support for the annotation of the human genome that is being carried out by the GENCODE as part of the ENCODE project.

In this work we carried out a comprehensive identification and validation of GENCODE annotations for the human genome. We have identified peptides covering 40% of the GENCODE genes (lets use 7 because we can), validated the translation to protein of "novel" and "putative" protein coding transcripts and confirmed the presence of multiple alternative gene products for 179 genes. This is the largest set of genes that have been reliably confirmed to generate multiple alternative gene products.

We were able to demonstrate that these genes were significantly enriched in alternative isoforms characterized by subtle differences in protein sequence, suggesting that the translation of alternative transcripts may be subject to selective constraints. These findings are backed up by parallel analyses of mouse and Drosophila proteomics experiments.

This work highlights the growing importance of proteomics in validating predicted proteins, and also as a complement to large-scale annotation efforts such as the ENCODE project.

Quest for phylogenetically-stable gene markers.

Salvador Capella-Gutierrez[1], Frank Kauff[2] and Toni Gabaldón[1]

[1]Bioinformatics and Genomics Programme. Centre for Genomic Regulation (CRG) and UPF. Barcelona, Spain.
[2]Molecular Phylogenetics. Dept. of Biology. University of Kaiserslautern. Kaiserslautern, Germany

Evolutionary relationships among bacterial species have been traditionally inferred using ribosomal genes, especially 16S, given their high ubiquitousness and degree of conservation. With the increased availability of completely sequenced genomes, however, we have now a whole range of genes at our disposal. This though poses a new question: which genes are most informative to establish the relationships of the organisms being studied. The prevalence of Horizontal Gene Transfer (HGT) in bacteria imposes the need to distinguish among genes likely reflecting the underlying species relationships from those displaying alternative scenarios. Ideally, an informative set of genes should be present in the studied species and remain informative when more taxa are added to the study . The later condition is critical to identify the most suitable markers to be obtained from unsequenced species, a necessary step in targeted studies aiming for broad taxonomic sampling. Here we present a method to identify small subsets of genes that recover the same species tree that we obtain from a larger, complete set of widespread single-copy genes. The method divides available genomes into two groups, the first group is used to identify subset of genes that recover with high support the species tree inferred with all genes, while the second group is used to validate the selection. To validate our method, we decided to use a set of 62 cyanobacterial genomes, these species are important due to their possible role as producers of alternative sources of energy. 43 of these genomes were used as part of the first group while 19 of them were part of the validation process. After performing the first stage of our method, we got a total of 252 widespread single-copy gene sets and three subsets of candidates with 7, 10 and 22 genes. During the second stage, we validate the initial selection of the 7 genes as the golden set of genes that are phylogenetically stable across all cyanobacterial genomes used so far. This genes are currently being used in a larger project aiming to uncover the cyanobacterial tree of life.

Bioinformatics and Molecular Studies the: Glyceralehyde-3-Phosphate Dehydrogenase of Pseudomonas Syringae

Elkhalfi B.[1], Serrano A.[2], Soukri A.[1]

[1] Laboratory of Physiology & Genetics Molecular. Faculty of Sciences, Ain Chock, Casablanca, Morocco.
[2] Institute IBVF (CSIC Univ.of Seville), Seville, Spain
Email: bouchra.elkhalfi@gmail.com

Pseudomonas syringae pv. Tomato DC300, the causal organism of bacterial speck, common disease of tomato, which's the mode of infection, is understood, according to the Advanced Molecular Biology, Genomics and Proteomics, Pseudomonas syringae produces a number of proteins that promote infection and draw nutrients in the tomato plant. Among these proteins glyceralehyde-3-phosphate dehydrogenase (GAPDH).

With Bioinformatics tools, we searched the genomic and peptide sequences of the Pseudomonas syringae GAPDH, we have aligned the three genes of GAPDH and we have designated specific primers for Ral-Time PCR to analyze the expression of mRNA.

On the other hand, we designated others primers to amplify the three genes by PCR, but before with bioinformatics programs we predicted the ideal conditions for this PCR.

Following the genomic study we found that Pseudomonas syringae pv. Tomato has three GAPDH with different size and function and which probably at least be involved in the infection. This was confirmed also by the expression of mRNA. The fragments amplified, will be cloned and expressed to analyze and to produce specific antibodies to check the catalytic and / or physiological functions from each of the three GAPDH proteins of this pathogenic bacterium.

Keywords: GAPDHs, Pseudomonas syringae, Bioinformatic, Real-Time PCR, Pathogenicity.

Hardwired clusters displayed by tree-child networks

Gabriel Cardona[1], Mercè Llabrés[1] and Francesc Rosselló[1]

[1]Dept. Mathematics and Computer Science, Univ. of the Balearic Islands, 07122 Palma (Spain)

Phylogenetic networks are models of evolutionary histories that allow for the representation of reticulate evolutionary events like recombinations, hybridizations, or horizontal gene transfers, where one species does not derive directly from a single species, but from the interaction of several (usually, two) species [1]. Formally, a *phylogenetic network* on a set of taxa S is a rooted directed acyclic graph with its leaves bijectively labeled in S. The nodes in a phylogenetic network can be classified into *tree nodes*, with indegree ≤ 1, and *hybrid nodes*, with indegree ≥ 2. Tree nodes represent species, and hybrid nodes represent reticulate evolutionary events. We shall assume that each hybrid node has outdegree 1, its only child representing the byresult of the reticulate event. A phylogenetic network is *tree-child* when every internal node has a child of tree type [1, Sect. 6.11.4]. Tree-child (TC) networks include phylogenetic trees, galled trees and normal networks, and they have been proposed by S. Willson as the class where meaningful phylogenetic networks should be searched.

The *cluster* of a tree node v in a phylogenetic network N is the set of all taxa corresponding to descendant leaves of v. The set of clusters of all tree nodes of a network N is called its set of *hardwired clusters*, $C_{hard}(N)$. A non-redundant (that is, not containing an arc (u,v) and another path $u \rightarrow v$) network having a given set C of hardwired clusters can be built using the construction of the cluster network of C as explained in [1, Sect. 6.4], basically the Hasse diagram of C with the containment relation. But this network need not be unique, and this cluster network need not be minimal in any sense among all networks N such that $C_{hard}(N)=C$.

In this work we investigate the hardwired representation of clusters by means of TC networks. In particular:

- We prove that $C= C_{hard}(N)$ for some TC network N iff for every $X \in C$ there exists $l \in X$ such that for every $X' \in C$, if $l \in X'$, then X and X' are one contained into the other. This generalizes the usual compatibility characterization of clusters of phylogenetic trees.
- We prove that if $C=C_{hard}(N)$ for some TC network iff the cluster network built from C is TC
- We prove that if $C=C_{hard}(N)$ for some TC network, then the cluster network built from C is the unique non-redundant TC network with this property, and moreover it is minimal (in the sense that it has the least number of nodes and the least sum of indegrees of hybrid nodes) among all networks N' such that $C_{hard}(N')=C$.

[1] D. Huson, R. Rupp, C. Scornavacca. *Phylogenetic Networks. CUP (2011)*

A new imbalance index for phylogenetic trees

Arnau Mir[1], Francesc Rosselló[1], and Lucía Rotger[1]

[1]*Dept. Mathematics and Computer Science, Univ. of the Balearic Islands, 07122 Palma (Spain)*

One of the most thoroughly studied properties of rooted phylogenetic trees is their degree of imbalance, that is, the degree to which sister internal nodes tend to have the same number of descendant taxa. Several indices that measure this degree of imbalance have been introduced so far in the literature, including Colless's index I for resolved trees (the sum over all internal nodes of the absolute value of the difference between the number of descendant leaves of their children) [1] or Sackin's index S for arbitrary (i.e., not necessarily resolved) trees (the sum over all leaves of their depths) [2].

In this work we propose and study a new imbalance index F for phylogenetic trees, which we call the *total cophenetic value*: the sum, over all pairs of different leaves, of the depth of their least common ancestor. This index makes sense for arbitrary trees and has more resolutive power than both Colless's and Sackin's indices. We prove that:

- The tree with n leaves with the largest F value is the caterpillar
- The tree with n leaves with the smallest F value is the star tree
- The resolved tree with n leaves with the smallest F value is the *most balanced tree with n leaves*: the only (up to relabelling of the leaves) rooted tree where each internal node has children with numbers of descendant leaves that differ in at most 1. It is interesting to point out that this tree has also the smallest Colless's and Sackin's indices, but for those indices there may exist many other non-balanced trees yielding this minimum while for our index it is the only tree yielding this minimum.
- In particular, we obtain the range of possible values of F on binary trees with n leaves, which turns out to be of the order of one power of n larger than the ranges for I and S: the small size of the ranges of the latter is one of their drawbacks [3].
- We have obtained an explicit formula for the distribution F_n of F on binary trees with n leaves, as well as for the expected value of F_n, under the uniform model (where all trees with the same number of leaves have the same probability). In particular, we prove that $E(F_n)\sim n^4/24$ as $n\to\infty$. It is known that the expected values of both I_n and S_n are $\pi^{1/2}n^3$ [4].
- We have computed the F index of a sample of 2000 binary trees contained in the TreeBase and we have compared the empirical distributions for different numbers of leaves (n=3,…,91) with the theoretical distribution mentioned in the previous bullet. They turn out to be similar, which could be considered as a further evidence that real phylogenetic trees are consistent with the uniform distribution [5].

[1] D. H. Colless. *Syst. Zool.* **31** (1982), 100-104
[2] K. Sao, R. Sokal. *Syst. Zool.* **39** (1990), 266-276
[3] J. S. Rogers. *Syst. Biol.* **45** (1996), 99-110
[4] M. Blum et al. *Ann. Appl. Prob.* **16** (2006), 2195-2214
[5] A. McKenzie, M. Steel. *Math. Biosc.* **170** (2001), 91-112.

STRUCTURAL FEATURES OF FUNGAL RIBOSOMAL DNA

Marco Marconi, Sara Gago, Ana Alastruey-Izquierdo, María José Buitrago, Juan Luis Rodrígez-Tudela, Manuel Cuenca-Estrella, Isabel Cuesta.

Mycology Reference Laboratory, National Centre of Microbiology, Instituto de Salud Carlos III, Carretera Majadahonda-Pozuelo km 2, Majadahonda 28220, Madrid (Spain).

Ribosomal DNA (rDNA) array has been thoroughly studied in both, prokaryotes and eukaryotes, including fungi. This region consists of tandem repeats of the ribosomal RNA (rRNA) precursor sequence separated by an intergenic spacer region (IGS). The fungal rDNA internal transcribed spacer (ITS) is the genetic region that has been used as a phylogenetic marker, as it contains enough variability to discriminate at species level, therefore it has been reported as the gold standard for molecular identification of fungal species. Moreover, the use of a correct antifungal therapy to treat Invasive Fungal Infection (IFI) is species-dependent and there is a direct relationship between the time it takes to administer an appropriate treatment and patient mortality, very high in this type of infection (40%-90%).

However, the taxonomic value of ITS region has been questioned, as it does not provide a sufficient level of discrimination in some genera (i.e. *Aspergillus*, *Cryptococcus*, *Fusarium*). This raises the need to identify a target with a better phylogenetic resolution than the one used so far and that keeps the same multi copy nature necessary to obtain a good sensitivity during the amplification by PCR. In recent years, the IGS has been explored for typing at species and strain levels. However, this region has not been sufficiently studied or exploited (unlike the ITS regions) due to its large size and the high number of repeat motifs. In order to know if the IGS is a phylogenetic marker suitable for fungal species identification or genotyping, the variability within tandem-repeats ribosomal DNA arrays have been studied. These sequences have been extracted by blasting from the traces of genomic sequencing projects of two species of yeast, *Candida albicans and Candida tropicalis* available in public databases. The ribosomal DNA copy number in each genome and, the partial single nucleotide polymorphisms (transitions and transversions) within the rDNA array have been identified. Analysis of the data reveals that there are not variation within ITS region repeats, supporting its use as a phylogenetic marker. On the other hand, the IGS region of the rDNA accumulates most of the variation of the rDNA array, being this variability in *C. albicans* concentrated in IGS1. The characteristics of the IGS sequence suggest the presence of microsatellites which can be useful in the molecular typing of these fungi. Further studies with higher number of strains for each species are warranted to confirm these results.

Horizontal gene transfer in fungal species.

Marina Marcet-Houben[1] and Toni Gabaldón[1]

[1]*Centre for Genomic Regulation (CRG) and UPF, Dr, Aiguader, 88, 08003 Barcelona, Spain*

Horizontal gene transfer (HGT) has been a key factor in the evolution of prokaryotes but its existence within the evolution of eukaryotic species is not well studied. Fungi are one of the most well sampled groups within eukaryotes in terms of number of completely sequenced genomes, providing an excellent resource in which to apply comparative genomics. Additionally, few of them have been scanned for the presence of HGT events. The few species that have been previously scanned for HGT events have shown that while not as abundant as in prokaryotes, HGT events can be detected both, between prokaryotes and fungi and between fungal species.

In order to provide a larger overview on the scope of horizontal gene transfer within fungal genomes we searched 60 complete genomes for HGT events using phylogenomic tools. Briefly we scanned the completely sequenced genomes for proteins with a particular phylogenetic distribution. These proteins had to be present in a low number of fungal species, they could not have homologs in other eukaryotic species and should be widely dispersed in prokaryotic organisms. We found 713 proteins dispersed in the 60 fungal species that complied with these thresholds.

The proteins were grouped into families. A maximum likelihood phylogenetic tree including the prokaryotic homologs was reconstructed for each family. The phylogenetic trees enabled us to separate the transferred protein families into 235 monophyletic HGT events. Phylogenies were also used to transfer the annotations from the prokaryotic proteins to the fungal transferred proteins, detect a possible prokaryote donor and to locate in which node in the fungal species tree each event occurred. Surprisingly the mapping clearly showed that the occurrence of HGT events in fungi varies between the different fungal groups. Pezizomycotina, a large group of filamentous fungi, showed a greater tendency of transferring and retaining proteins from prokaryotes.

While the number of HGT events for each fungal species does not approach the levels found in prokaryotes, the effect of these events on the evolution of fungi may be significant. For instance we detected the transfer of a prokaryotic arsenate reductase protein to two fungal species. This gene is essential to confer resistance to arsenic to these two species.

Additionally, our analysis was limited to clear-cut cases of transferences of prokaryotic genes to fungal genomes. Recently there have been several studies showing that HGT events between fungal species are also important. These transferences are not limited to single genes but also include transferences of clusters of genes and even of complete genomes. This leads us to believe the great importance of HGT events in the evolution of fungi and the need to get a deep understanding on the reach of such important evolutionary events.

M. Marcet-Houben and T. Gabaldón. *Trends Genet* ***26(1)*** (2011) 5-8.

Recombination detection and its use in inferring the recent human evolutionary history

Marta Melé[1‡], Marc Pybus[1], Asif Javed[2], Laxmi Parida[2], Francesc Calafell[1,], Jaume Bertranpetit[1], The Genographic Consortium
1 IBE, Institute of Evolutionary Biology (UPF-CSIC), CEXS-UPF-PRBB, Barcelona, Catalonia, Spain, **2** Computational Biology Center, IBM T J Watson Research, Yorktown, New York, United States of America,

Recombination is one of the main forces shaping genome diversity, but the information it generates is often overlooked. A recombination event creates a junction between two parental sequences that may be transmitted to the subsequent generations. Just like mutations, these junctions carry evidence of the shared past of the sequences and could potentially be used as genetic markers.

Based in a combinatoric algorithm [1], we developed a program, IRiS [2,3], aimed at detecting past recombination events from extant sequences and also aimed at finding the breakpoint location and which are the recombinant sequences. We validated and calibrated IRiS for the human genome using coalescent simulations replicating standard human demographic history and a variable recombination rate model, and we fine-tuned IRiS parameters to simultaneously optimize for false discovery rate, sensitivity, and accuracy in placing the recombination events in the sequence. Thus in the same way that SNP data along a chromosome constitutes a haplotype, recombination data makes a recotype, which can be inferred for specific chromosomes and pooled for populations, describing, in a new approach, the recombinational landscape.

Then, we applied IRiS to data collected over 30 populations in the Old World corresponding to 1240 males which were genotyped in 1250 SNPs on the X chromosome with the aim of using the patterns of recombination to make inferences on the history of these populations [4]. Specifically, based on the number of recombinations detected in each population we were able to infer their effective population size. We have found that Sub-Saharan African populations have an N_e that is approximately 4 times greater than those of non-African populations and that outside of Africa, South Asian populations had the largest N_e. We also observe that the patterns of recombinational diversity of these populations correlate with distance out of Africa if that distance is measured along a path crossing South Arabia. No such correlation is found through a Sinai route, suggesting that anatomically modern humans first left Africa through the Bab-el-Mandeb strait rather than through present Egypt.

The present work opens a new paradigm in recombination studies as individual events may be detected through a computational approach and may be used in very different genetic and genomic approaches.

1) Estimating the ancestral recombinations graph (ARG) as compatible networks of SNP patterns. Parida L, **Melé M**, Calafell F, Bertranpetit J; Genographic Consortium.
J Comput Biol. 2008 Nov;15(9):1133-54. PMID: 18844583

2) A new method to reconstruct recombination events at a genomic scale. **Melé M**, Javed A, Pybus M, Calafell F, Parida L, Bertranpetit J; Genographic Consortium Members.
PLoS Comput Biol. 2010 Nov 24;6(11):e1001010. PMID: 21124860.

3) IRiS: construction of ARG networks at genomic scales. Javed A, Pybus M, **Melé M**, Utro F, Bertranpetit J, Calafell F, Parida L. Bioinformatics. 2011 Sep 1;27(17):2448-50. Epub 2011 Jul 15.PMID: 21765095

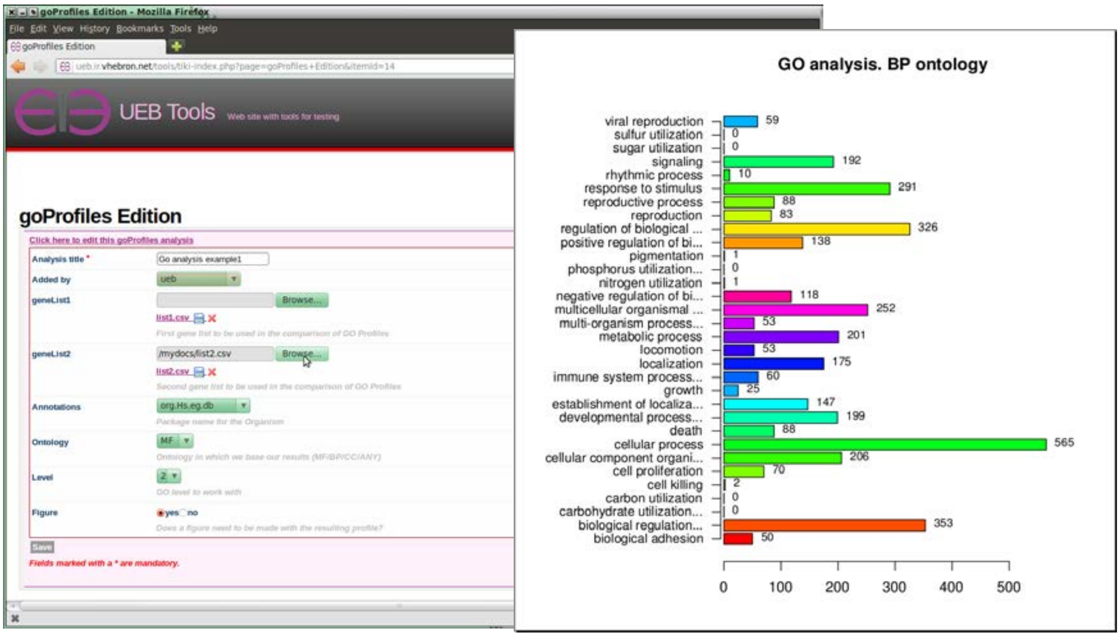
4) Recombination Gives a New Insight in the Effective Population Size and the History of the Old World Human Populations. **Melé M**, Javed A, Pybus M, Zalloua P, Haber M, Comas D, Netea MG, Balanovsky O, Balanovska E, Jin L, Yang Y, Pitchappan RM, Arunkumar G, Parida L, Calafell F, Bertranpetit J; The Genographic Consortium. Mol Biol Evol. 2011 Oct 10. [Epub ahead of print] PMID: 21890475.

Deploying a tools website in a Bioinformatics unit with Tiki & PluginR

Ferran Briansó[1], Alex Sánchez-Pla[1,2], and Xavier de Pedro[1]

[1]*Unitat d'Estadística i Bioinformàtica. Vall d'Hebron Institut de Recerca*
[2]*Departament d'Estadística. Facultat de Biologia. Universitat de Barcelona*

There have been some recent advances in web-based graphical user interfaces (GUI) dedicated to an easy and efficient exploitation of R scripts [1]. The use of Tiki Wiki CMS Groupware [2] (also known as ”Tiki", which stands for “Tightly Integrated Knowledge Infrastructure“) in combination with PluginR [3] has proved to be an excellent approach to implement these kind of tools with special applicability in the Bioinformatics area, as presented in the JBI 2010 edition [4]. The relatively new PluginR for Tiki, has so far allowed the development of GUIs that, easily and safely, use R scripts in different areas, such as research on the Teaching and Learning field [5] or Microarray Analyses Pipelines for biomedical scientists [6]. This communication seeks to point out the main advantages and disadvantages found up to date with the use of this versatile Tiki+PluginR technology, within the frame of a Bioinformatics unit that provides support services to a biomedical research center. It will be shown the work done in developing some web-based tools (currently available at http://ueb.ir.vhebron.net/tools), such as a tailored easy-to-use heat map generator or a new (and still under construction) tool for the comparison of gene lists using the Bioconductor goProfiles package [7].



goProfiles Edition

Click here to edit this goProfiles analysis

Analysis title * Go analysis example1

Added by ueb

geneList1 list1.csv
First gene list to be used in the comparison of GO Profiles

geneList2 /mydocs/list2.csv
list2.csv
Second gene list to be used in the comparison of GO Profiles

Annotations org.Hs.eg.db
Package name for the Organism

Ontology MF
Ontology in which we base our results (MF/BP/CC/ANY)

Level 2
GO level to work with

Figure ● yes ○ no
Does a figure need to be made with the resulting profile?

Fields marked with a * are mandatory.

Screenshots of the new **goProfiles** web-based tool

[1] Newton, R; Wernisch, L. *R News* **7:2** (2007) 32-35.
[2] Tiki Wiki CMS Groupware: Software made the wiki way (2011) - http://tiki.org
[3] Documentation, how to setup & use Tiki: PluginR (2011) - http://doc.tiki.org/PluginR
[4] de Pedro, X; Sánchez-Pla, A. In *X Jornadas de Bioinformática*, Málaga (2010).
[5] de Pedro et al. In *VI International Congress of University Teaching and Innovation, Barcelona* (2010).
[6] de Pedro, X; Sánchez-Pla, A. In *19th Annual International Conference on Intelligent Systems for Molecular Biology and 10th European Conference on Computational Biology, Vienna* (2011).
[7] Salicrú et al. *BMC Bioinformatics* 12:401 (2011).

Two Component Systems:
Physiological effect of a third component

Rui Alves[1], Baldiri Salvado[1], Ester Vilaprinyo[2], Albert Sorribas[1], and Rui Alves[1]

[1]IRBLleida & Universitat de Lleida
[2]Fundació IMIM

Signal transduction systems mediate the response and adaptation of organisms to environmental changes. In prokaryotes, this signal transduction is often done through Two Component Systems (TCS). These TCS are phosphotransfer protein cascades, and in their prototypical form they are composed by a kinase that senses the environmental signals (SK) and by a response regulator (RR) that regulates the cellular response. This basic motif can be modified by the addition of a third protein that interacts either with the SK or the RR in a way that could change the dynamic response of the TCS module.

This work aims at understanding the effect of such an additional protein (which we call “third component”) on the functional properties of a prototypical TCS. To do so we build mathematical models of TCS with alternative designs for their interaction with that third component. These mathematical models are analyzed in order to identify the differences in dynamic behavior inherent to each design, with respect to functionally relevant properties such as sensitivity to changes in either the parameter values or the molecular concentrations, temporal responsiveness, possibility of multiple steady states, or stochastic fluctuations in the system. The differences are then correlated to the physiological requirements that impinge on the functioning of the TCS. This analysis sheds light on both, the dynamic behavior of synthetically designed TCS, and the conditions under which natural selection might favor each of the designs.

We find that a third component that modulates SK activity increases the parameter space where a bistable response of the TCS module to signals is possible, if SK is monofunctional, but decreases it when the SK is bifunctional. The presence of a third component that modulates RR activity decreases the parameter space where a bistable response of the TCS module to signals is possible.

Is Cloud Computing an attractive alternative for Bioinformatics?

Torreño O.[1], Ramet D.[1], Karlsson T.J.M[1], Lago J[2]., Bodenhofer, U.[3] and Trelles O[1].

1Computer Architecture Department, Málaga University, Spain
2Fundacion IAVANTE, Consejería de Salud, Junta de Andalucía
3Institute of Bioinformatics, Johannes Kepler University, Linz, Austria

Cloud computing (CC), defined as a model for enabling convenient, on-demand network access to a shared pool of configurable computing resources with minimal management effort or service provider interaction, is rapidly growing in importance as life science organizations are deluged with data from multiple sources. Limited funding and budgets make it difficult for many organizations to build the infrastructure necessary to tackle these challenges. Cloud computing appears for many as a promising alternative to in-house expansion.

However, CC is still in its early days, and most life science organizations are still proceeding cautiously to test its feasibility and determine which applications run best in that mode. Yet driven by continual acceleration in the rates of data generation and the desire for processor-intensive applications, these organizations continue to increase their cloud utilization and the diversity of applications they run there.

In this document we present a detailed study to assess the comparative performance of CC against several other multiprocessor architectures for the fast porting of sequential legacy applications to be run in parallel. We have designed a Map-Reduce like procedure named Mr.Cirrus (Map-Reduce High Level Clouds) [1] to benchmark the different architectures under a unified approach avoiding the inclusion of additional latency or overheads associated with the software implementation.

Shared and distributed memory multiprocessor architectures have been tested, including clusters of commodity PC against the most important Clouds environments such as Microsoft-Azure [2]; IBM-SmartCloud [3], and Amazon EC2 [4]. Exhaustive tests have been performed on well-known bioinformatics applications with diverse computational patterns and different I-O work loads. These applications meet some of the following criteria: (1) covers a broad range of computational patterns; (2) represent legacy applications to evaluate the portability of legacy software; (3) they target big data-sets or are CPU-demanding.

Our initial results show:

- Functionality: Azure offers a PaaS (Platform as a service) model with a high level of abstraction, which means less control over the hardware. Amazon offers a IaaS (Infrastructure as a Service) model, so basically once kind of hardware is selected it becomes accessible throughout the network. Users have administration control over the virtualizated machine. IBM offers the two previous options and also a SaaS (Software as a Service) model, that is basically the sotfware produced by IBM for desktop machines running in the cloud. These CC models are associated with the functionality offered by the CC provider and the capabililties for software designers to deploy their parallel models.
- Up-to 30% on speed-up differences are observed among the different implementations in the different platforms.
- Strong dependencies on I-O workload. Special care must be observed when process big data sets. Even simple strategies such as compress-send-uncompress data could produce important saving on computational costs.
- Trade-off between specific code editing versus unmodified legacy code ruining in parallel: although a good collection of legacy software is available in bioinformatics, writing specific parallel software still results profitable inversion.
- Start-up: documentation, easiness to start implementation, functionality. Learning time depends both on the complexity of the API and in the available documentation. CC do not scape from this observation. Differences in the slope of the learning curve are observed associated with the profusion of examples, forums, developer environments, etc.

We expect that the results from this study could lead to new or adapted HPC tools and assist in the making decision process for the most appropriated cloud architecture to solve specific bioinformatics problems.

Acknowledgments

This work has been partially supported by grant AT2009-0025 "Acciones Integradas-MICINN", and FP7-VENUS.C projects. OT and DR are funded by RIRAAF-ISCIII proyect. OTS and JK are members of the GN5-National Institute of Bioinformtics, Spain

[1] Daniel Ramet, Juan Lago, Johan Karlsson, Juan Falgueras y Oswaldo Trelles; (2011), "Mr-Cirrus: Implementación de Map-Reduce bajo MPI para la ejecución paralela de programas secuenciales; XXII Jornadas de Paralelismo; Las Palmas de Gran Canaria, España
[2] What is the Windows Azure platform?; http://www.microsoft.com/windowsazure/
[3] IBM smart Cloud: http://www.ibm.com/cloud-computing/us/en/
[4] Amazon Elastic Compute Cloud: http://aws.amazon.com/es/ec2/

Cell DB, a complete integrative biological database and *RESTful* web service API

Ignacio Medina[1,3], Alejandro de María[1], Marta Bleda[2], Luz Garcia[1], Joaquín Dopazo[1,2,3]

[1]Department of Bioinformatics and Genomics, Centro de Investigación Príncipe Felipe (CIPF), Valencia, Spain
[2]CIBER de Enfermedades Raras (CIBERER), Valencia, Spain
[3]Functional Genomics Node (INB), CIPF, Valencia, Spain

During the last years, because of advances in high-throughput technologies, we have witnessed an unprecedented growth of repositories and databases storing relevant biological data. Today we have more biological information than ever but unfortunately the current status of many these repositories is not what a researcher would like most of the time. Some of the most common problems are: a) information is spread out in hundreds of small repositories and databases, b) lack of standards between different repositories, c) unmaintained databases, d) specific and unconnected information, … All these problems make very difficult: a) to integrate or join many different sources into only one database to work or analyze experiments; b) to access and query this information in programmatically way.

To solve all these problems we have developed a relational database to contain the most relevant biological information about genomic features and proteins, gene expression regulation, functional annotation, genomic variation and systems biology information. About one hundred tables have been well designed and normalized into one database. We have integrated the most relevant repositories such as Ensembl, Uniprot, IntAct, Reactome, ...The information integrated covers:

- Core features: genes, transcripts, exons, proteins, ...
- Regulatory: TFBS, miRNA, CTCF, ...
- Functional annotation: OBO ontologies (Gene ontology, desease ontology), ...
- Genomic variation: dbSNP, HapMap, 1000Genome project, COSMIC, ...
- Systems biology: IntAct , Reactome, gene co-expression, ...

To make all this database accessible to researchers, an exhaustive *RESTful* web service *API* has been implemented. This API contains more than 100 methods that will facilitate researchers to query and obtain different biological information from a single database. Another benefit is that researchers can make queries about different biological topics and link all this information together.

Multiscale modeling of different segment peptides of Hemagglutinin from influenza viruses

Jorge M. Antunes[1], Bruno L. Victor[1] and Cláudio M. Soares[1]

1 Instituto de Tecnologia Química e Biológica, Universidade Nova de Lisboa. Portugal

Influenza viruses (IV) are responsible for worldwide outbreaks of flu, which can have dramatic consequences, as demonstrated by the recent pandemic caused by the new strain of IV, H1N1. Unfortunately, due to high genetic variability and insufficient knowledge about the infection process, we have very few preventing weapons against these viruses. Hemagglutinin (HA) is one of the key infection regulators of the entire infection process enrolled by these organisms. This protein comprises two chains, HA1 and HA2. HA1 contains surface regions used in the initial step of the infection, to bind to cell surface receptors containing sialic acid. After internalization of the virus into endosomes, the low pH conditions induce dramatic structural changes in HA1. These changes will allow the fusion peptide (FP), found in the N-terminal of the HA2, to become exposed and in close contact with the endosome membrane bilayers. This proximity is essential because it will allow the FP to interact and promote its disruption resulting ultimately in the fusion of both the endosome and viral membrane bilayers. Another important peptide segment found in the C-terminal of the HA2 chain is the Transmembrane Peptide (TM). This segment peptide, besides being the anchor of the HA protein to the membranes of the virus, can also be important in the final steps of the fusion process. Due to the importance these two peptide segments have in the fusion process of IV, and because the relationship between their function and structure is still not completely understood, we have use a multiscale approach to address this problem.

In this work we present several Molecular Dynamics studies with the TM peptide in the presence of membrane bilayers. With these studies we performed an exhaustive structural characterization of the stability of the segment peptide in a DMPC membrane bilayer, and we have also identified and confirmed the importance of specific sequence requirements, by performing several mutational studies. To further complement the characterization of segment peptides from HA, we have also performed several studies with the FP, now performing a multiscale approach, using both coarse-grained and atomistic molecular dynamics simulations. These simulations aim at identifying the structural determinants of the stability and interaction of this peptide with lipid bilayers.

Understanding and predicting adverse drug reactions from biological and chemical spaces

Miquel Duran-Frigola[1] and Patrick Aloy[1,2]

[1]Institute for Research in Biomedicine. Joint IRB-BSC Programme in Computational Biology
[2]Institució Catalana de Recerca i Estudis Avançats (ICREA)

In pharmacology, it is key to identify the mechanisms of drug action in order to understand and infer undesired side effects. Here, we see side effects as a phenotypic signal in response to drug treatment, and we aim at bridging the gap between such responses and the molecular basis of drug action. To achieve so, we navigate both the chemical and the biological spaces seeking for regions that markedly correlate with adverse drug reactions (ADRs).

Given a collection of ADRs and the drugs reported to cause them, we gather (a) their structural features and (b) the proteins they interact with. We then perform an enrichment analysis of these annotations to eventually point out biological and chemical features that are associated to each side effect. In consequence, the method provides us with a set of ADR predictors that are well defined and easy to interpret.

Models of regulatory genomics

Sonja Althammer[1], Amadis Pages[1], Eduardo Eyras[1,2]

[1] *Universitat Pompeu Fabra. Dr. Aiguader 88, E08003, Barcelona, Spain*
[2] *Catalan Institution for Research and Advanced Studies (ICREA), Passeig Lluís Companys 23, E08010, Barcelona, Spain*

Background
High-throughput sequencing (HTS) has trigger much progress recently in the study of transcription regulation. In particular, it has provided new possibilities to describe the expression regulation of genes in terms of the ChIP-Seq signal for histone modifications.

Results
We show that using HTS signal for histone marks in only one condition present several confounding effects. We describe a novel approach based on the density changes between two different conditions, which mitigate those effects. In this way, we can describe expression changes under two conditions using epigenetic changes between the same conditions. We used our tool Pyicos to calculate epigenetic changes using ENCODE data from different types of experiments: ChIP-Seq data for several histone modifications, Bisulfite-Seq data for DNA methylations, DNase-Seq data for open chromatin regions and finally RNA-Seq data for the expression of transcripts. Using a machine learning approach, we are able to build a predictive model of the expression changes between the cell-lines K562 and Gm12878 with high accuracy. Additionally, testing this model on a different pair of cell-lines, Hsmm and Hmec, we achieve a comparable accuracy. Furthermore, we estimated the predictive value of the different types of epigenetic elements in several genomic regions (promoter, first intron, etc.) using feature selection techniques. We found that the most informative features to predict the expression changes depend on the promoter properties of the gene and on its exon-intron structure. Finally, we obtained the minimal set of regulatory features that is sufficient to predict expression changes with high enough accuracy.

Conclusions
We are able to generate a model based on epigenetic changes between two cell-lines that is generic enough to predict the expression change between any pair of cell-lines. Our model can capture the importance of the relative location of epigenetic elements to get insight into the regulation of expression between two cell-lines. Moreover, our results indicate that there is a generic histone code of expression regulation that may depend of the genomic context of the gene.

The Collaborative Writing of “Ten Simple Rules for Getting Help from Online Communities”: are Open Collaborative Papers the Future of Science?

Dall’Olio Giovanni Marco[1], Invergo Brandon[1], Bertranpetit Jaume[1] and Laayouni Hafid[1]

[1] IBE, Institut de Biologia Evolutiva, CEXS-UPF (Barcelona, Spain)

In March 2011, our lab tried a different approach toward the process of writing a scientific article. We originally wanted to write a paper for the PLoS Computational Biology “Ten Simple Rules” series, dedicated to how to interact with online scientific communities. However, as the topic of the article was rather general and of interest for a broad audience, we decided that the only effective way to achieve a “community consensus” in the paper was to involve as many people as possible in it, by searching for new collaborations from Internet, and making the whole process of writing the manuscript public. This has lead to our first experience of “open collaborative academic paper”.

The first step has been to write a preliminary draft of the document, and publish it on a publicly accessible wiki, WikiGenes. Then, we sent messages to several mailing lists and online scientific communities, as well as on twitter and other resources, inviting people to contribute. After few hours, we started getting contributions from researchers who we had never met before, but who in the following weeks participated to the manuscript and made critical contributions to it. After six months, the Ten Simple Rules paper was published on PLoS Computational Biology[1], featuring eight new authors that we did not know personally before submitting the open call, but who greatly increased the value of the manuscript.

One of the most important advantages of open collaborative writing is that this type of articles are more likely to meet a community consensus criteria, eliminating possible impartialities and merging together multiple points of view in the same manuscript. Given the increasing complexity of modern scientific research, and the availability of big datasets in the public databases, which will make the problem of multiple interpretations of the data more frequent, we think that Open Collaborative Papers may have a place in the future of science. Thus, in this talk we will describe our experience in organizing the writing of a Open Collaborative Paper, and give our insight of what went well and what went wrong in our case. We will give some suggestions for future similar projects, and discuss possible issues that may arise from this model.

References:

[1] Dall’Olio, G.M. et al., 2011. Ten Simple Rules for Getting Help from Online Scientific Communities P. E. Bourne, ed. *PLoS Computational Biology*, 7(9), p.e1002202. Available at: http://dx.plos.org/10.1371/journal.pcbi.1002202 [Accessed September 29, 2011].

MOWSERV 2: the multi-repository version of the INB integrated client

Alfredo Martinez[1] ,Oswaldo Trelles[1]

1 Computer Architecture Department, University of Malaga, 29071 Málaga, Spain

{amatrinezl,ots}uma.es

Abstract.

Background On the recent years the Bioinformatics tools growth on diversity and heterogeneous has been more than a fact. The variety of these tools is one of the keys of data analysis flexibility and potential but, also, carries some disadvantages such as un-standardized service interfaces, different data formats for inputs and outputs, different technologies or ways to access those recourses, etc. As a consequence, applications aimed to make an integrate way to use and exploit those tools have become outdated for nowadays needs.

Results We have developed a new version of the current platform offered by the Spanish National Institute of Bioinformatics (INB), MOWServ [1]. MOWServ was developed in 2005 and since that new technological advances and new community needs have been raised. The new client has been developed with the objective of solve all those lacks from its predecessor. The following table (Table 1) shows a comparative between both versions:

	MOWServ	MOWServ2
Technology	PHP as the main technology. Adobe Flex for some newer components like the object editor. CGI for some specific functionality (Web Service invocation).	Adobe Flex for the client side and Java for the Sever (MAPI [2] Web Services).
Interface	Versatile but not very intuitive. Limited usability as reported by users.	Desktop-like interface based on windows, icons and menus. Double-click invocation of services in conjunction with advanced execution-control, extensibility of viewer plug-ins, plus a file-based browsing style and organization of favorite tools.
Repository	MOWServ can connect to any BioMoby-based repository. Each installation of MOWServ is configured to connect to one repository.	Supports multiple repositories. Each user can switch between several repositories on real time.
Service discovering	Implemented by Magallanes [3]. The service and recourses discovering is limited to the repository used by the installation of MOWServ.	Implemented also by Magallanes. Search into each repository that exploits MOWServ2.
Extension capabilities	Each new functionality is coded and integrated in the application code. No development framework supplied.	Offers a development framework which supports the creation of new components/functionalities by third parties.

Table 1. Comparative between both clients.

Discussions As an approach toward usability and intuitive interfaces, Bioinformatics applications and tool integration are mandatory. Mixing integration and a highly flexible platform MOWServ2 client has been developed, providing a configurable platform and an easy way to extend functionality by developing new components.

Acknowledgements This work has been partially financed by the National Institute for Bioinformatics (www.inab.org) a platform of Genoma-España and the ACGT-EU project "Advancing Clinico-Genomic Trials on Cancer" (EU-contract no.026996). The authors would like to thank the Bitlab research team for invaluable help.

References

[1] Ramirez S, Muñoz-Mérida A, Karlsson J, García M, Perez-Pulido A., Claros G., Trelles O: MOWServ: a web client for integration of bioinformatic resources; Nucleic Acid Research, 2010 1-6. DOI 10.1093/nar/gkq497

[2] Sergio Ramirez, Johan Karlsson, Oswaldo Trelles; "MAPI: towards the integrated exploitation of bioinformatics Web Services"; BMC Bioinformatics 2011, Vol 12:419.

[3] Ríos J., Karlsson J and Trelles O.: Magallanes: a Web Services discovery and automatic workflow composition tool; BMC Bioinformatics, 10:334. (2009)

An integrative approach to the detection of cancer driver genes

Abel Gonzalez-Perez[1] and Nuria Lopez-Bigas[1,2]

[1]*Research Group on Biomedical Informatics-GRIB, Universitat Pompeu Fabra, Barcelona Biomedical Research Park-PRBB, Barcelona, Spain*
[2]*Institució Catalana de Recerca i Estudis Avançats (ICREA), Barcelona, Spain*

Cancer genome sequencing projects [1,2] are uncovering large lists of alterations across cohorts of samples of several tumor types. Although these efforts have opened up the possibility to identify which specific genomic and epigenomic alterations are essential for the development of each tumor type, this issue is far from resolved. Some approaches have been taken in the past couple of years to uncover such cancer *driver* alterations and the genes that bear them in a few tumor types. Typically, such approaches rely either on a) an evaluation of the potential impact of genomic alterations on protein expression or function [3,4] or b) an assessment of the statistical over-representation of a given gene within the alterations list of a tumor genome [5,6]. Recently, this second type of –recurrence based– strategies have been used in conjunction with other expected features of driver genes, such as functional closeness and mutual exclusivity [7,8].

Nevertheless, these recurrence based approaches normally fail to take into account differences in background alterations rate between samples. Moreover they are unable to detect lowly recurrently altered genes which may be important to tumorigenesis. On the other hand, functional impact approaches give no information whatsoever on the frequency of alterations and therefore are prone to underestimate marginally deleterious mutations with a potential for driverness.

We have approached the problem of detecting potential driver genes from large cohorts of tumor genome samples by combining four different sources of information. First, we evaluate the functional impact of non-synonymous SNVs (nsSNVs) on genes by combining three predictive methods using a variation of an idea recently developed in our group [9] and identify those that tend to accumulate functional nsSNVs. Second, we identify which genes are targets of nsSNVs and/or Copy Number Variations (CNVs) more frequently than expected by chance. We also use the enrichement for functional connections between likely driver genes (in the previous list) and knowledge available from at least one previous independent study. We demonstrate our integrative approach using the catalog of genomic alterations identified in cohorts of samples of glioblastoma multiforme and serous ovarian carcinoma.

[1] The Cancer Genome Atlas. *Nature* **455** (2008).
[2] International network of cancer genome projects. *Nature* **464** (2010) 993-998.
[3] Reva B et al. *Nucleic Acids Research* (2011).
[4] Carter H et al. *Cancer Research* **69** (2009) 6660-6667.
[5] Greenman C et al. *Nature* **446** (2007) 153-158.
[6] Akavia UD et al. *Cell* **143** (2010) 1005-1017
[7] Cerami E et al. *PLoS ONE* **5** (2010).
[8] Ciriello G et al. *Genome Research* (2011).
[9] Gonzalez-Perez A et al. *AJHG* **88** (2011) 440-449.

Poster presentations (even numbers)

Day 1. **Even numbered posters**
Wednesday 25th 9:00-19:00

SNAPE: A MODEL FOR DETECTING SNVS AND COMPUTING IDENTICAL BY DESCENT REGIONS.

EMANUELE RAINERI[1], LUCA FERRETTI[2], FRANCESC CASTRO-GINER[1], AND SIMON HEATH[1]

Abstract. SNV detection is subject to uncertainty which comes from different sources, of which at least two are clear: there are mapping ambiguities due to repetitive regions in the DNA, and there are sequencing errors. It is necessary to reflect this uncertainty in the SNV calling to avoid mistakes in further analysis layers. It is also useful to keep in consideration heuristics which have been tested in practice, *e.g.* that a nucleotide is to be included in the pileup only if it is observed in both forward and reverse reads. After inferring SNVs, a logical following step is to analyze the differences between SNVs in two samples. Here we present a software (SNAPE) which computes the probability of every possible genotype at every position of a sequenced region, and the probability that two samples have the same genotype at a certain position. Furthermore, we show how this model can be used to derive the coordinates of IBD regions in pedigrees where at least some of the members have been sequenced.

[1]Centro Nacional de Análisis Genómico (CNAG),08028 Barcelona, Spain
E-mail address: `emanuele.raineri@gmail.com`

[2]Centre for Research in Agricultural Genomics (CRAG), 08193 Bellaterra, Spain

RUbioSeq: An automatic workflow for variant detection in NGS studies.

Miriam Rubio Camarillo[1], Gonzalo Gómez-López[2], David G. Pisano[2]

[1]*Structural Computational Biology Group, Structural Biology and BioComputing Programme, Spanish National Cancer Research Centre (CNIO). Madrid, Spain.*

[2]*Bioinformatics Unit, Structural Biology and BioComputing Programme, Spanish National Cancer Research Centre (CNIO). Madrid, Spain.*

Recent advances in sequencing technology present novel opportunities for many applications in life sciences. The vast number of data produced by next-generation sequencing (NGS) techniques poses significant computational challenges and many computational steps are required to translate this output into high-quality results. Many of these bioinformatic analyses consist on a set of stages that are repetitive and routinely executed. Developing a workflow able to perform all stages in an automatic and reliable way is crucial to eliminate manual steps and to speed up result generation.

Here, we present a variant calling workflow executed in a HPC platform, which automatically executes all involved stages in this process using state-of-art software. Briefly, our process includes (a) short-read alignment with a combination of BWA[1] and BFAST[2] aligners, (b) a GATK-based variant calling protocol for *indels* and SNPs[3] (c) SNP effect predictor[4] to annotate the variants. Our program accepts raw data in Fastq format from Illumina platform and generates a result file with the variants and its biological impact prediction ranked by detection quality score. Parallel multiple sample execution has been developed in order to reduce the processing time. Quality and errors check points for input data and files created during the execution are also performed.

In order to include future improvements, the software has been designed in a modular structure to permit easy adaptation and extension. Reliability of the program has been already proved through experimental assays which have validated the supplied variants. This workflow has been established by the Bioinformatics Unit at CNIO to perform their variant calling analyses. Future implementations include MethylC-seq and ChIP-seq data analyses.

[1] Li H. and Durbin R. Bioinformatics. Vol. 26 (2010), pages 589–595.
[2] Nils Homer .PlosOne. Vol. 4 (2009). Issue 11. e7767.
[3] Aaron McKenna. Genome Research. Vol. 20 (2010), pages1297–1303
[4] William McLaren. Bioinformatics. Vol. 26 (2010) , pages 2069–2070.

Evaluation of single CpG sites as proxies of CpG island methylation states at the genome scale.

Víctor Barrera[1], Miguel A. Peinado[1]

[1] *Institut de Medicina Predictiva i Personalitzada del Càncer (IMPPC), Badalona; Barcelona, Spain.*

The DNA methylation state of a CpG island is a faithful marker of activity for the associated gene. Since the methylation profile of CpG islands, especially in silenced genes, is relatively homogenous, different techniques used in the routine analysis in CpG island methylation are often limited to the analysis of one or a few CpG sites within the region of interest. This is adequate if the reporter information of the selected site (or sites) has been previously verified by analysis of larger DNA stretches within the interrogated CpG island. Recent massive sequencing analyses have offered, for the first time, high resolution maps of DNA methylation at the genome scale in a handful of human samples. These genomic approaches are still unfeasible for most experimental designs which must rely on the application of alternative strategies in large scale DNA methylation studies. To investigate the representativeness of individual CpG sites we have analyzed the methylation correlation between selected CpGs and the corresponding CpG island using data from Lister et al[1]. Due to its use in different large scale approaches, we have chosen the CpG located within the SmaI and the HpaII restriction sites as surrogate indicators used by differnt experimental approaches. The global and specific accuracy of SmaI and HpaII sites in regard to the methylation states of the corresponding CpG island are investigated and analyzed according to different genomic features. Beyond the global validation of extrapolating the state of a single site as surrogate marker of the whole CpG island, this study may raise new insights into the functional significance of discordant sites.

[1] Lister R et al. *Nature* **462** (2009) 315-322.

A framework for the analysis of transcriptional and post-transcriptional gene regulatory networks

Marta Bleda[1], Ignacio Medina[1,2], and Joaquín Dopazo[1,2,3]

[1]CIBER de Enfermedades Raras (CIBERER), Valencia, Spain
[2]Department of Bioinformatics and Genomics, Centro de Investigación Príncipe Felipe (CIPF), Valencia, Spain
[3]Functional Genomics Node (INB), CIPF, Valencia, Spain

Transcription factors (TFs) and microRNAs (miRNAs) are important regulators in the control of gene expression, permitting the normal development of cells. Alterations in these elements have been extensively related with human malignancies including cancer. So far, the study of gene regulatory networks has shed some light over the control of gene expression, however, an analysis tool integrating transcriptional and post-transcriptional information does not exist yet.

We have developed an algorithm able to identify the regulatory implications of a set of deregulated genes. This implementation is supported by an inclusive database of transcriptional and post-transcriptional regulatory information based, mainly, on the action of transcription factors and microRNAs. Previous knowledge of the relationship between diseases and the deregulation of this elements is also included. Interestingly, all this information is programmatically accessible through a RESTful API web services.

This tool ease the exploration of regulatory networks enabling a better understanding of functional modularity and network integrity under specific perturbations. Results are represented graphically through a web user-friendly interface. In addition, expression data can also be integrated to better analyze the regulatory cause of perturbations. The characterization of its topological features has enabled a better understanding of functional modularity and network integrity under specific perturbations.

A genomic signature for endospore formation

Ana Abecasis[1], Mónica Serrano[2], Renato J. Alves[3], Leonor Quintais[4], Adriano O. Henriques[2], Jose B. Pereira-Leal[3]

[1]Instituto de Higiene e Medicina Tropical - Universidade Nova de Lisboa
[2]Instituto de Tecnologia Química e Biológica - Universidade Nova de Lisboa
[3]Instituto Gulbenkian de Ciência
[4]European Bioinformatics Institute

Bacterial endospores are the most resistant cell type known to man, able to withstand extremes of temperature, pressure, chemical injury and time. They are of particular interest as the endospore is the infective particle in a variety of human and livestock diseases. Endosporulation, the process of forming the endospore, is characterized by the morphogenesis of a specialized dormant and highly resistant endospore within a mother cell. We analyzed the evolutionary profile of the regulators of the endosporulation program, as well as its machinery to identify a conserved core of ~100 genes that we propose represent the minimal machinery for endosporulation. Our results also indicate that endosporulation was likely invented once, at the base of the Firmicutes phylum (c.a., 3 Billion years ago), that it is unrelated to other bacterial spore differentiation programs, and that it involved the invention of new genes and functions, as well as the co-option of ancestral, house keeping functions. Based on this knowledge we defined a genomic signature able to distinguish endospore forming organisms based on complete genome sequences, that we show to be robust against phylogenetic proximity and other artifacts. This signature includes previously uncharacterized genes, that we now show by site-directed mutagenesis in Bacillus subtilis to be essential for sporulation, and to be cell-type specific, thus further validating this genomic signature. We predict that a series of extremophylic organisms, as well as many unexpected gut bacteria will be able to form endospores. Our work paves the way to new drug targeting strategies that prevent the formation of the endospore, rather than the bacteriostatic or bactericidal activities of current antibiotics that cannot affect the highly resistant bacterial endospore.

The Three-Dimensional Architecture of a Bacterial Genome

Umbarger MA[1], Toro E[2], Wright MA[1], Porreca GJ[1], Baù D[3], Hong S-H[2,4], Fero HJ[2], Marti-Renom MA[3], McAdams HH[2], Shapiro L[2], Dekker J[5], and Church G[1].

1 Department of Genetics, Harvard Medical School, Boston, Massachusetts 02115, USA.

2 Department of Developmental Biology, Stanford University School of Medicine, Beckman Center, B300, Stanford, CA 94305

3 Structural Genomics Laboratory, Centro de Investigación Príncipe Felipe, 46012 Valencia, Spain.

4 Department of Physics, Stanford University School of Humanities and Sciences, Stanford, CA 94305

5 Program in Gene Function and Expression, Department of Biochemistry and Molecular Pharmacology, University of Massachusetts Medical School, Worcester MA 01605, USA

We have determined the three-dimensional (3D) architecture of the Caulobacter crescentus genome by combining genome-wide chromatin interaction detection, live-cell imaging, and computational modeling. Using chromosome conformation capture carboncopy (5C) technology, we derive ~13 Kb resolution 3D models of the Caulobacter genome. These models illustrate that the genome is ellipsoidal with periodically arranged arms. The parS sites, a pair of short contiguous sequence elements involved in chromosome segregation, are positioned at one pole of this structure, where they nucleate a compact chromatin conformation. Both 5C and imaging experiments demonstrate that placing these sequence elements at new genomic positions yields large-scale rotations of the genome within the cell. Utilizing automated fluorescent imaging, we orient the genome within the cell and illustrate that within the resolution of our data the parS proximal region is the only portion of the genome stably attached to the cell envelope. Our approach provides an experimental paradigm for deriving insight into the cis-determinants of 3D genome architecture..

Quantification of miRNA-mRNA interactions

Ander Muniategui[1], Rubén Nogales-Cadenas[2], Miguél Vázquez[3], Xabier L. Aranguren[4], Xabier Agirre[5], Aernout Luttun[4], Felipe Prosper[5], Alberto Pascual-Montano[2] and Angel Rubio[1]

[1]*Group of Bioinformatics, CEIT and TECNUN, University of Navarra, San Sebastian, Spain,*
[2]*National Center for Biotechnology-CSIC, Madrid, Spain,*
[3]*Structural Biology and Biocomputing Programme, Spanish National Cancer Research Centre (CNIO), Madrid, Spain,*
[4]*Center for Molecular and Vascular Biology, Katholieke Universiteit Leuven, B-3000 Leuven, Belgium*
[5]*Hematology Department and Area of Cell Therapy, Clínica Universidad de Navarra, Foundation for Applied Medical Research, University of Navarra, Pamplona, Spain*

miRNAs are small RNA molecules ('22*nt*) that interact with their mRNA targets inhibiting their translation into proteins and cleaving the target mRNA. This second effect diminishes the overall expression of the target mRNA. This deregulation is key in a wide range of biological processes and human diseases . The complexity of miRNA regulatory mechanism in mammals makes experimental validation and computational prediction of targets challenging [2]. Still, the number of experimentally-validated targets is very small and the expected amounts of false positives on sequence based predictions are too high [3].

In order to achieve more accurate relationships, some methods that combine expression data of miRNAs and mRNAs with sequence based predictions have emerged [4-8]. They assume that the expression values of miRNAs and their targets have to be inversely related due to down-regulation effects from miRNAs. For instance, GenMiR++ [4] uses a Bayesian framework to infer the probability of a putative interaction of being real from expression data. HOCTAR uses correlation values to decipher the targets of intronic miRNAs from the expression values of their host genes [5]. MAGIA uses the mutual information [6]. Jayaswal et al. [7] cluster each expression data and search significant interactions by using t-test. Finally, Li et al. [8] apply Partial Least Squares Regression.

In our work, we use LASSO regression with non-positive constraints (Talasso). We enforce the sparseness of the solution and restrict the predicted mRNA targets to those with down-regulation effects from miRNAs. We have compared TaLasso, GenMiR++ and Pearson correlation to two datasets with matched samples with expression from mRNA and miRNAs.

We compared them by measuring their ability to retrieve experimentally-validated targets and analysing the biological relevancy of the predicted interactions. The top ranked interactions recovered by TaLasso are especially enriched in experimentally validated targets and the functions of the genes with mRNA transcripts in the top-ranked interactions are meaningful. This is not the case of GenMiR++ and correlation. Furthermore, adding non-positive constraints improves the specificity of the prediction using LASSO regression. TaLasso is available as Matlab or R code and as a web-based tool for human miRNAs at http://talasso.cnb.csic.es/.

References

[1] Alvarez-Garcia I, et al. *Development* **132** (2005) 4653.
[2] Alexiou P et al. *Bioinformatics* **25** (2008) 3049.
[3] Yue D et al. Curr Genomics 10 (2009) 478.
[4] Huang JC et al. *Nat Methods* **4** (2007) 1045.
[5] Gennarino VA et al. *Genome Res* **19** (2009) 481.
[6] Sales G et al. *Nucleic Acids Res* **38** (2010) W352.
[7] Jayaswal V et al. *BMC Genomics* **12** (2011) 138.
[8] Li X et al. *BMC Med Genomics* **4** (2011).

Drafting a large genome at high quality: Multi-platform sequence assembly and integration with genetic and physical maps of sugar beet (Beta vulgaris)

Juliane C. Dohm[1,2], André E. Minoche[1,2], Daniela Holtgräwe[3], Thomas Rosleff Sörensen[3], Richard Reinhardt[4], Hans Lehrach[1], Bernd Weisshaar[3], Heinz Himmelbauer[2]

[1]Max Planck Institute for Molecular Genetics, Berlin, Germany
[2]Centre of Genomic Regulation (CRG) and UPF, Ultrasequencing Unit, Barcelona, Spain
[3]Bielefeld University, CeBiTec, Bielefeld, Germany
[4]Max Planck Institute for Plant Breeding Research, Max Planck Genome Centre Cologne, Cologne, Germany

With second generation sequencing the production of large data amounts is now feasible, but building high-quality de novo drafts remains a challenge, especially for large and repeat-rich genomes. To ensure the success of such a genome project several resources are necessary like large-insert genomic libraries as well as genetic and physical mapping information.

The genome of sugar beet (Beta vulgaris) has an estimated size of 758 Mbp in nine chromosomes and consists of about 65% repetitive sequence. We sequenced a doubled haploid accession in a whole genome shotgun approach with data generated on Roche/454, Solexa/Illumina, and Sanger platforms. The single (21x coverage) and paired read data (369x library coverage) were assembled into scaffolds using the Newbler software. Subsequent integration with genetic and physical maps of sugar beet substantially improved the scaffold size. Of the resulting 596 Mbp of sequence data two-thirds could be assigned to chromosomes, and each chromosome is covered on average by 25 large scaffolds. The total N50 size is 1.53 Mbp with 87 sequences covering 50% of the assembly. The consensus sequence was inspected for insertions and deletions by mapping 93x coverage of Illumina read data which were not included in the assembly previously. After consensus correction we performed evidence-based gene prediction using 583 million mRNA-seq reads resulting in about 32,000 genes supported by evidence.

CpGcluster2: An improved clustering algorithm to predict CpG islands

Michael Hackenberg, José L. Oliver, Guillermo Barturen, Antonio Rueda and Francisco Dios

Dpto. de Genética, Facultad de Ciencias, Universidad de Granada, Campus de Fuentenueva s/n, 18071-Granada, Spain & Lab. de Bioinformática, Centro de Investigación Biomédica, Campus de la Salud, Avda. del Conocimiento s/n, 18100-Granada, Spain

The methylation of CpG dinucleotides is an important epigenetic modification, required in mammals for embryonic development, genomic imprinting and X-chromosome inactivation. Unmethylated stretches of clustered CpG dinucleotides (CpG islands or CGIs) are conventionally detected by sliding window approaches using a large parameter space formed by the thresholds of length, CpG fraction and G+C content. On the contrary, CpGcluster [1] is based on the physical distance between consecutive CpGs on the chromosome, being able to predict directly clusters of CpGs. By assigning a *p*-value, the most statistically significant CpG clusters can be predicted as CGIs. We introduce here a new version of this algorithm (CpGcluster2: http://bioinfo2.ugr.es/CpGcluster/), which incorporates several important improvements: 1) the only search parameter (the distance between two consecutive CpGs), normally specified as a percentile of the distance distribution, is now tabulated for a great variety of genome assemblies, which improves CGI prediction in sequences below 10 kb; 2) genome and chromosome intersection (the point separating the intra-cluster from the inter-cluster distances) can be now automatically detected; 3) the CpG probability can now be provided by the user which is important for short sequences in species that are not in our database; and 4) the CGI repository includes now the precomputed predictions for all the animal genome assemblies available at the UCSC server, providing also direct links to the UCSC Genome Browser and NGSmethDB [2]. These links allow to visualize the predicted CGIs in a wide genome context, as well as to inspect their methylation profiles in different tissues. These features make CpGcluster2 the algorithm of choice for epigenome studies on CGIs.

[1] M. Hackenberg et al. *BMC Bioinformatics* **7** (2006) 446.
[2] M. Hackenberg et al. *Nucleic Acids Res.* **39** (2011) D75-9.

On the Universal Random-Like Structure of Genomes

Hernán Dopazo[1], Becher Verónica[2], and François Serra[1]

[1]Evolutionary Genomics Laboratory. Bioinformatics and Genomics Department. Centro de Investigación Príncipe Felipe. Valencia. Spain
[2]Computation Department. Facultad de Ciencias Exactas y Naturales. Universidad de Buenos Aires. Argentina

Most biologists agree that the genomes of complex organisms require a higher number of genetic directives, hence, larger genomes [1]. Lynch and Conery hypothesized that genome size and its complexity inevitable increased as the consequence of population size reduction in linages [2]. Definitions of biological complexity have not been completely successful because they work well just for specific properties [3]. In contrast, formalizations of complexity at the genome level have received fewer objections [4,5]. These studies however, focused on specific species, and used methods that involve genome fragmentation; hence, those results are local to the analyzed DNA fragments. Here we ask for the global combinatorial structure of DNA in biodiversity. Is there a common law in the DNA sequences of all genomes?

In this work we tested the hypothesis that there is a common combinatorial structure of DNA in all genomes. Our hypothesis is that there is a random-like structure of DNA along all diversity of life. To test it, we define a complexity measure based on a classical method used in data compression [6] and applicable to arbitrarily large sequences introducing no fragmentation. The method detects regularities due to repeats of any length, at any distance, and other structural correlations.

As the main result we report that the ratio of genome complexity to size remained almost maximal and unchanged along six orders of magnitude in genome size, covering all biological diversity. We observe a uniform complexity increases with genome size for phages, bacteria, unicellular eukaryotes, fungi, plants, and animals. Major deviations from maximal genome complexity correspond to polyploid species. Diploidization [7] -the process by which a polyploid genome turns into a diploid one- guaranties the return to almost maximum complexity. We formulate two general hypotheses: 1- almost maximal combinatorial structure of DNA sequence is a common characteristic of genomes throughout biological diversity; 2- increases in the combinatorial complexity of DNA only occur by mechanisms of genome amplification, and subsequent accumulation of DNA sequence mutations, transpositions and/or deletions of genetic material. Our hypothesis can be falsified if a single recent polyploid genome with a random-like DNA structure is found; or if a non-polyploid genome shows a non-random DNA structure.

[1] Maynard Smith J, Szathmáry E. The major transitions in evolution. New York: W. H. Freman. (1995)

[2] Lynch M, Conery JS The origins of genome complexity. *Science* **302** (2003) 1401–1404.

[3] McShea D. Perspective: Metazoan complexity and evolution: Is there a trend? *Evolution* **50** (1996) 477–492.

[4] Azbel M Universality in a DNA statistical structure. *Phys Rev Lett* **75** (1995) 168–171.

[5] Adami C, Ofria C, Collier TC Evolution of biological complexity. *Proc Natl Acad Sci USA* **97** (2000) 4463–4468.

[6] Adjeroh D, Bell T, Mukherjee A. The Burrows-Wheeler transform: data compression, suffix arrays, and pattern matching. Springer **41** (2008) 21–24.

[7] Wolfe KH. Yesterday's polyploids and the mystery of diploidization. *Nat Rev Genet* **2** (2001) 333–341.

A large-scale survey reveals that chromosomal copy-number alterations significantly affect gene modules involved in cancer initiation and progression

Eva Alloza[1], Fátima Al-Shahrour[1,2], Juan C Cigudosa[3,4] and Joaquín Dopazo[1,3,5]

1 Department of Bioinformatics and Genomics, Centro de Investigación Príncipe Felipe (CIPF), Valencia, Spain. 2 Broad Institute, 7 Cambridge Center, Cambridge, MA 02142, USA. 3 CIBER de Enfermedades Raras (CIBERER), ISCIII, CIPF, Valencia, Spain. 4 Molecular Cytogenetics Group. Centro Nacional de Investigaciones Oncologicas (CNIO), Madrid, Spain. 5 Functional Genomics Node (INB), CIPF, Valencia, Spain.

Background: Recent observations point towards the existence of a large number of neighborhoods composed of functionally-related gene modules that lie together in the genome. This local component in the distribution of the functionality across chromosomes is probably affecting the own chromosomal architecture by limiting the possibilities in which genes can be arranged and distributed across the genome. As a direct consequence of this fact it is therefore presumable that diseases such as cancer, harboring DNA copy number alterations (CNAs), will have a symptomatology strongly dependent on modules of functionally-related genes rather than on a unique "important" gene.

Methods: We carried out a systematic analysis of more than 140,000 observations of CNAs in cancers and searched by enrichments in gene functional modules associated to high frequencies of loss or gains.

Results: The analysis of CNAs in cancers clearly demonstrates the existence of a significant pattern of loss of gene modules functionally related to cancer initiation and progression along with the amplification of modules of genes related to unspecific defense against xenobiotics (probably chemotherapeutical agents). With the extension of this analysis to an Array-CGH dataset (glioblastomas) from The Cancer Genome Atlas we demonstrate the validity of this approach to investigate the functional impact of CNAs.

Conclusions: The presented results indicate promising clinical and therapeutic implications. Our findings also directly point out to the necessity of adopting a function-centric, rather a gene-centric, view in the understanding of phenotypes or diseases harboring CNAs.

Protein disorder in the Centrosome correlates with complexity in cell types number.

G. S. Nido[1], R. Méndez[1], A. Pascual-García[1], D. Abia[1], U. Bastolla[1], and H. Gomes-Dos Santos[1]

1 Centro de Biología Molecular "Severo Ochoa", (CSIC-UAM), Cantoblanco, 28049 Madrid, Spain.

The aim of this work was to study the properties and the evolution of proteins that constitute the Centrosome, the complex molecular assembly that regulates the division and differentiation of animal cells. We found that centrosomal proteins are predicted to be significantly enriched in disordered and coiled-coil regions, more phosphorylated and longer than control proteins of the same organism. Interestingly, the ratio of these properties in centrosomal and control proteins tends to increase with the number of cell-types. We reconstructed indels evolution, finding that indels significantly increase disorder in both centrosomal and control proteins, at a rate that is typically larger along branches associated with a large growth in cell-types number, and larger for centrosomal than for control proteins. Substitutions show a similar trend for coiled-coil, but they contribute less to the evolution of disorder. Our results suggest that the increase in cell-types number in animal evolution is correlated with the gain of disordered and coiled-coil regions in centrosomal proteins, establishing a connection between organism and molecular complexity. We argue that the structural plasticity conferred to the Centrosome by disordered regions and phosphorylation plays an important role in its mechanical properties and its regulation in space and time[1].

We are extending this study to a higher number of representative species. These organisms have been selected from OMA database[2] (http://omabrowser.org), including unicellular eukaryotes, fungi, plants and animals as well as a subset of parasites. This protein database relates proteins considering orthologue relationships at different levels.
Preliminary results are in agreement with the conclusions exposed about the relation between disordered regions and the cell-types number.

[1] G. S. Nido et al. Molecular. BioSystems, vol 8 (2012), 353–367.
[2] A. Schneider et al. Bioinformatics, Vol. 23 (2007), 2180–2182.

Pyicos: A versatile toolkit for the analysis of high-throughput sequencing data

Sonja Althammer[1], Juan González-Vallinas[1], Cecilia Ballaré[2], Miguel Beato[1,2], and Eduardo Eyras[1,3]

1 Universitat Pompeu Fabra. Dr. Aiguader 88, E08003, Barcelona, Spain. 2 Centre for Genomic Regulation (CRG). Dr. Aiguader 88, E08003 Barcelona, Spain. 3 Catalan Institution for Research and Advanced Studies (ICREA), Passeig Lluís Companys 23, E08010, Barcelona, Spain.

Motivation: High-Throughput Sequencing (HTS) has revolutionized gene regulation studies and is now fundamental for the detection of protein-DNA and protein-RNA binding, as well as for measuring RNA expression. With increasing variety and sequencing-depth of HTS datasets, the need for more flexible and memory- efficient tools to analyse them is growing.

Results: We describe *Pyicos*, a powerful toolkit for the analysis of mapped reads from diverse HTS experiments*:* ChIP-Seq, either punctuated or broad signals, CLIP-Seq, and RNA-Seq. We prove the effectiveness of *Pyicos* to select for significant signals and show that its accuracy is comparable and sometimes superior to that of methods specifically designed for each particular type of experiment. *Pyicos* facilitates the analysis of a variety of HTS datatypes through its flexibility and memory efficiency, providing a useful framework for data integration into models of regulatory genomics.

Availability: Open source software, with tutorials and protocol files, is available at http://regulatorygenomics.upf.edu/pyicos or as a Galaxy server at http://regulatorygenomics.upf.edu/galaxy

Efficient alignment of pyrosequencing reads for re-sequencing applications

Francisco Fernandes[1,2], Paulo GS da Fonseca[1], Luis MS Russo[1,2], Arlindo L Oliveira[1,2] and Ana T Freitas[1,2]

1 Instituto de Engenharia de Sistemas e Computadores: Investigação e Desenvolvimento (INESC-ID), R. Alves Redol 9, 1000-029 Lisboa, Portugal. 2 Instituto Superior Técnico-Universidade Técnica de Lisboa (IST/UTL), Av. Rovisco Pais, 1049-001 Lisboa, Portugal.

Background: Over the past few years, new massively parallel DNA sequencing technologies have emerged. These platforms generate massive amounts of data per run, greatly reducing the cost of DNA sequencing. However, these techniques also raise important computational difficulties mostly due to the huge volume of data produced, but also because of some of their specific characteristics such as read length and sequencing errors. Among the most critical problems is that of efficiently and accurately mapping reads to a reference genome in the context of re-sequencing projects.

Results: We present an efficient method for the local alignment of pyrosequencing reads produced by the GS FLX (454) system against a reference sequence. Our approach explores the characteristics of the data in these re- sequencing applications and uses state of the art indexing techniques combined with a flexible seed-based approach, leading to a fast and accurate algorithm which needs very little user parameterization. An evaluation performed using real and simulated data shows that our proposed method outperforms a number of mainstream tools on the quantity and quality of successful alignments, as well as on the execution time.

Conclusions: The proposed methodology was implemented in a software tool called TAPyR–Tool for the Alignment of Pyrosequencing Reads–which is publicly available from http://www.tapyr.net.

nucleR: a package for non-parametric nucleosome positioning

Oscar Flores[1] and Modesto Orozco[1,2,3]

1 IRB-BSC Joint Research Program on Computational Biology, Institute of Research in Biomedicine, Baldiri i Reixac 10. 2 Department of Biochemistry and Molecular Biology, University of Barcelona, Avinguda Diagonal 645. 3 Instituto Nacional de Bioinformática. Parc Científic de Barcelona. Baldiri i Reixac 10, Barcelona 08028, Spain.

Summary: nucleR is an R/Bioconductor package for a flexible and fast recognition of nucleosome positioning from next generation sequencing and tiling arrays experiments. The software is integrated with standard high-throughput genomics R packages and allows for in situ visualization as well as to export results to common genome browser formats.

Availability: Additional information and methodological details can be found at http://mmb.pcb.ub.es/nucleR

MAPI: towards the integrated exploitation of bioinformatics Web Services

S. Ramirez[1], J. Karlsson[1], O. Trelles[1]

[1] *Department of Computer Architecture, Málaga University, Spain*

Data analysis in bioinformatics is increasingly becoming distributed in the form of Web Services (WS) because of the strong requirements for data storage and CPU time. However, access to such WS has not been standardized in the bioinformatics community, something which has resulted in a proliferation of tools, whose dispersion and heterogeneity complicate the integrated exploitation of available data processing capacity. In this scenario, the requirement for standardized and user-friendly tool interfaces to support end-users is clear. Unfortunately, software client developers need to spend a lot of time to handle/transform different data formats and WS protocols.

In order to facilitate the construction of software clients and facilitate the integrated use of this variety of tools, we present the Modular Application Programming Interface (MAPI) [1] which provides the necessary functionality for a uniform representation of metadata for WS and repositories. The software framework provides support for metadata management and service execution.

MAPI has been designed to adapt to the requirements of client software due to a modular organization which means that only the modules needed in a specific client must be installed. The module functionality can be extended with access to new types of services and metadata repositories without the need for re-writing the software client.

In [1], we demonstrated the utility and versatility of the software library by the implementation of several clients, covering different aspects of integrated data processing; ranging from service discovery to service invocation with advanced features such as workflows composition and asynchronous services calls to multiple types of Web Services including those registered in repositories (e.g. GRID-based, SOAP, BioMOBY, R-Bioconductor and others).

Paper available at: http://www.biomedcentral.com/1471-2105/12/419

Acknowledgments

This work has been partially supported by grant AT2009-0025 "Acciones Integradas-MICINN" and FP7-VENUS-C projects. OTS and JK are members of the GN5-National Institute of Bioinformatics, Spain.

[1] Sergio Ramirez, Johan Karlsson and Oswaldo Trelles, “MAPI: towards the integrated exploitation of bioinformatics Web Services”, BMC Bioinformatics 2011, 12:419. doi:10.1186/1471-2105-12-419

firestar – advances in functional residue prediction

Paolo Maietta[1], Jose Manuel Rodríguez[2], Gonzalo Lopez[3], Alfonso Valencia[1,2] and Michael Tress.[1]

[1]Structural and Computational Biology Programme, Spanish National Cancer Research Centre (CNIO), Madrid, Spain
[2]National Institute of Biotechnology (INB), Spanish National Cancer Research Centre (CNIO), Madrid, Spain
[3]Center for Computational Biology and Bioinformatics, Columbia University, New York, NY 10032, USA

Here we present the recent developments in *firestar* [1], an expert system for predicting catalytic site residues and biologically relevant small molecule ligand binding residues. Several new features have been incorporated into *firestar* to improve the quality and coverage of the predictions. The scoring scheme has been improved and functional residues in the FireDB [2] repository are now classified in terms of their biological relevance, using evolutionary information, structural data and lists of known cognate ligands. This allows users to avoid false positives caused by crystallization conditions.

Prediction coverage has also been greatly improved with the extension of the FireDB database and by implementing HHblits, a variant of the highly successful HHsearch method [3], as an additional, more powerful means of detecting remotely related templates and generating alignments.

Many of the changes have been motivated by the critical assessment of techniques for protein structure prediction (CASP) ligand-binding prediction experiment [4], which provided us with a framework to test *firestar*. The server has been benchmarked against the CASP7 [5], CASP8 [4] and CASP9 [6] ligand binding prediction targets and was able to detect ligand binding residues for 94% of the sites over the three CASP editions. During CASP8 *firestar* outperformed all officially participating groups, and the recent improvements have improved *firestar*'s performance by 15% over the same targets.

Many improvements have also been made to the usability of the *firestar* server. The server predictions are now fully automatized and *firestar* has been implemented as a web service that is able to produce fast, high quality results in a high throughput mode.

Web site: http://appris.bioinfo.cnio.es/.

[1] G. Lopez et al. *NAR* ***39*** (2011) W235-41.
[2] G. Lopez et al. *NAR* ***35*** (2007) D219-23.
[3] J. Soeding. *Bioinformatics* ***21*** (2005) 951-60.
[4] G. Lopez et al. *Proteins* ***77*** (2009) 138-46.
[5] G. Lopez et al. *Proteins* ***77*** (2007) 165-74.
[6] T. Schmid et al. *Proteins* ***77*** (2011) 126-36.

ILOOPS server: A protein-protein interaction prediction utility based on local structural features

Manuel Marín-López[1], Jaume Bonet[1], Joan Planas-Iglesias[1] and Baldo Oliva[1].

[1]*Structural Bioinformatics Lab. GRIB. Universitat Pompeu Fabra*

Protein-protein interactions (PPIs) are crucial to understand how proteins perform their cellular functions. Therefore, correctly identifying the PPI network (or protein interactome) of a given organism is useful not only to shed light on the key molecular mechanisms behind a biological function but also to infer the functionality of a protein based on its interactions.

Recently, it has been suggested that several regions of the protein may be related to the molecular association, a central concept of the funnel-like intermolecular energy landscape used to describe PPIs [1]. Using BIANA [2], we extract all available Y2H PPIs as positive model of interaction. Conversely, negative data is obtained from Negatome [3]. We apply an association method to these data to extract characteristic signatures of both sets. These are used to score potential protein interacting pairs. Our results show that the combination of positive and negative signatures can be used to predict protein-protein interactions. Also, our data strongly suggests that it is the balance between typical interacting and non-interacting structural features in the protein surface which determine if a pair of proteins will interact or not.

ILOOPS is a web application for assessing protein-protein interaction based on those local structural features. The application takes a set of protein sequences, predicts the loops signatures and determines whether they interact or not. ILOOPS server also performs a homology modeling of the query proteins and assigns the loops to the modeled proteins. ILOOPS server is freely accessible on the web at http://sbi.imim.es/web/index.php/research/servers/iLoopsServer?page=index.php

[1] Wass, M.N. et al. *Molecular systems biology* **7** (2011) 469.
[2] Garcia-Garcia, J. et al. *BMC bioinformatics* **11** (2010) 56.
[3] Smialowski, P. et al. *Nucleic Acids Res*. **38** (2010) D540-4.

Support Vector Machines for Epitope Binding Classification

Fadi Chakik[1], Ahmad Shahin[1], and Walid Moudani[2]

[1]LaMA Group, Lebanese University, email: {fchakik, ashahin}@ul.edu.lb
[2]Lebanese University, email: wmoudani@ul.edu.lb

A key step in the development of an adaptive immune response to vaccines is the binding of peptides to molecules of the Major Histocompatibility Complex (MHC) for presentation to T lymphocytes, which are thereby activated. Several algorithms have been proposed for such binding predictions, but these are limited to a small number of MHC molecules and have imperfect prediction power. We are undertaking an exploration of the power gained by taking advantage of a natural representation of the protein sequence amino acid in terms of their composition, structural and a series of associated physicochemical properties [1] [2] to form a representative descriptor vectors. We are proposing to use dimensionality reduction techniques [3] [4] to preprocess the descriptor vectors before feeding them into well known statistical classifiers [5] [6] for binding prediction. In all cases, coupling dimensionality reduction techniques with the physicochemical properties leads to substantially higher values for our evaluation criteria (Area Under ROC Curve) which means that misclassification errors is reaching lower rates.

[1] J. R. Bock, and D. A. Gough, "Predicting protein—protein interactions from primary structure", Bioinformatics 2001, 17:455–60

[2] R. Karchin, K. Karplus, and D. Haussler, "Classifying G-protein coupled receptors with support vector machines". Bioinformatics 2002, 18:147–59.

[3] T. Balachander, R. Kothari, and H. Cualing, "An empirical comparison of dimensionality reduction techniques for pattern classification", In Proc. of the 7th International Conference on Artificial Neural Networks (ICANN 97), volume 1327, pages 589–594. Springer, Berlin, 1997.

[4] C. Shi and L. Chen. Feature dimension reduction for microarray data analysis using locally linear embedding. Proceedings of 3rd Asia-Pacific Bioinformatics Conference, 2005.

[5] V. Vapnik, (1998). Statistical Learning Theory. John Wiley & Sons, Inc., New York.

[6] B. Schölkopf, A. J. Smola, "Learning with kernels: support vector machines, regularization, optimization, and beyond", In Adaptive computation and machine learning Cambridge, Mass. MIT Press; 2002:xviii.

In silico optimization of the production of amino-acids in Escherichia coli

Rui Pereira[1], Paulo Vilaça[1,2], Miguel Rocha[2], Isabel Rocha[1]

[1]IBB-Institute for Biotechnology and Bioengineering / Centre of Biological Engineering, University of Minho, Campus de Gualtar, 4710-057 Braga, Portugal
[2] Department of Informatics / CCTC, University of Minho, Campus de Gualtar, 4710-057 Braga – PORTUGAL

The increasing need to replace chemical synthesis of compounds of interest by more environmentally friendly biological processes is driving the research for microbial cell factories. The industrial production of amino and organic acids includes several examples of success stories using microorganisms to convert inexpensive substrates into added value products. Traditionally, the design of such microbes relied on cycles of random mutagenesis followed by phenotypic selection [1], but a deeper knowledge of the microbial physiology allowed a more rational approach to this optimization problem [2,3]. However, this task is not straightforward, since the cell metabolism has proved to be highly complex and hard to predict.

One of the approaches to tackle this problem is to use Systems Biology simulation tools to predict the microorganism behavior when subjected to genetic modifications. Using genome scale stoichiometric models, such as the latest iAF1260 for *Escherichia coli* [4] one can simulate a great diversity of possible metabolic phenotypes under steady state conditions by imposing flux-balance constrains. The use of flux balanced analysis (FBA) allows the determination of flux values through all the reactions in the network under a set of environmental conditions and genetic manipulations, by using an objective function, such as the maximization of growth [5]. In this work, we used genetic algorithms, such as OptGene [6] to search for sets of gene knockouts that result in the overproduction *in silico* of amino-acids in *Escherichia coli.*

From all the proteinogenic amino-acids, glycine yielded the best results in the optimizations. A careful analysis of the *in silico* flux distribution in some of the mutants revealed an interesting and non-intuitive mechanism behind glycine accumulation. Furthermore, in these mutants the growth is coupled to the production of glycine, which makes them excellent candidates for *in vivo* implementation.

We are reaching a point where bioinformatics tools are advanced enough to aid in complex tasks, such as the optimization of microbial cell factories. Here we described an effort to optimize *in silico* the production of amino-acids in *Escherichia coli*, which resulted in the discovery of a potential set of knock-outs that leads to glycine overproduction. This serves to show the increasing importance of *in silico* optimizations to aid in the metabolic engineering projects, especially to search for non-intuitive beneficial genome modifications.

[1] K. Okamoto et al. *Bioscience, Biotechnology, and Biochemistry* **61** (1997) 1877-1882.
[2] A. Ozaki et al. *Agricultural and Biological Chemistry* **49** (1985) 2925–2930.
[3] M. Ikeda *Applied Microbiology and Biotechnology* **69** (2005) 615-626.
[4] A.M. Feist et al. *Metabolic Engineering* **12** (2010) 173-186.
[5] K.J. Kauffman et al. *Current Opinion in Biotechnology* **14** (2003) 491-496.
[6] K. Patil et al. *BMC Bioinformatics* **6** (2005) 308.

APIDConnector: exploring protein-interaction networks within Cytoscape

Diego Alonso-López, Alberto Risueño, Laura Bermúdez and Javier De Las Rivas

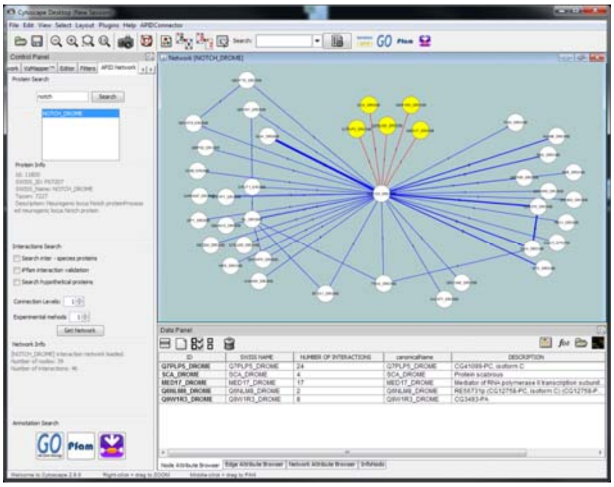
Bioinformatics, Centro de Investigación del Cáncer (CiC-IMBCC, CSIC/USAL) Salamanca, Spain

The advancement of genome and proteome-wide experimental technologies have introduced modern biology in the high complexity of living cells, where thousands of biomolecules work together in an interactive way, with many short and long range cross-talks and cross-regulations. To achieve a first level of understanding of such cellular complexity we need to unravel the physical interactions that occur between all the proteins that integrate an active working cell.

APID (Agile Protein Interaction DataAnalyzer) [1] is an interactive bioinformatic web-resource that has been developed to allow exploration and analysis of main currently known information about protein-protein interactions integrated and unified in a common and comparative platform. The analytical and integrative effort done in APID provides an open access frame where main experimentally validated protein-protein interactions deposit in six primary databases (*BIND*, *BioGRID*, *DIP*, *HPRD*, *IntAct* and *MINT*) are unified. The complete interactomes are rather complex systems so, despite that APID includes graphic displays to show sections of the interactome network, such graphical visualizations are quite limited when a specific study of a proteome subset or a protein family wants to be undertaken in detail. This is why we develop APIDConnector, a *Cytoscape* plugin that allows us to retrieve any query list of protein interactions stored in APID and integrate them into the *Cytoscape* environment. This tool provides a new way to build networks in Cytoscape that can be used in the last versions of Cytoscape (v 2.7 and 2.8).

With APIDConnector you can:

1. Search for a specific specie/protein and load its interaction network based on the APID query parameters.
2. Obtain a network seamlessly integrated into *Cytoscape*, which allows the use of all the Cytoscape functionality and third-party plugins to analyze and study the network.
3. Explore the proteins and interactions information of APID through nodes and edges attributes. Some of these attributes are used to create a custom visual style for the network view.
4. Search for *GO*, *Pfam* and *Interpro* annotations within the network and identify the nodes that apply for any of those annotations.



ID	SWISS NAME	NUMBER OF INTERACTIONS	canonicalName	DESCRIPTION
Q7PLP5_DROME	Q7PLP5_DROME	24	Q7PLP5_DROME	CG41099-PC, isoform C
SCA_DROME	SCA_DROME	4	SCA_DROME	Protein scabrous
MED17_DROME	MED17_DROME	17	MED17_DROME	Mediator of RNA polymerase II transcription subunit
Q9NLM8_DROME	Q9NLM8_DROME	2	Q9NLM8_DROME	RE56731p (CG12758-PC, isoform C) (CG12758-P...
Q9W1R3_DROME	Q9W1R3_DROME	8	Q9W1R3_DROME	CG3483-PA

Figure 1: Network visualization with APIDConnector

[1] Prieto, C. and De Las Rivas, J. (2006) APID: Agile Protein Interaction DataAnalyzer. Nucleic Acids Res., 34, W298–W302.

Exploiting protein flexibility to predict allosteric sites

Alejandro Panjkovich[1] and Xavier Daura[1,2]

[1]Institute of Biotechnology and Biomedicine (IBB), Universitat Autònoma de Barcelona (UAB), Bellaterra, E-08193, Spain
[2]Catalan Institution for Research and Advanced Studies (ICREA), Barcelona, E-08010, Spain

Protein allosteric sites are increasingly attracting the interest of medicinal chemists in the search for new types of targets and strategies to drug development. Given that allostery represents one of the most common and powerful means to regulate protein function, the traditional drug discovery approach of targeting active sites can be extended by targeting allosteric or regulatory protein pockets that may allow the discovery of not only novel drug-like inhibitors, but activators as well [1].

Moreover, allosteric sites present additional characteristics, such as modulable activity and less evolutionary pressure, which may facilitate the development of highly-specific allosteric drugs that can prevent side effects and readily complement traditional therapeutics.

Continuing with our previous work on allosteric sites from a structural and evolutionary perspective [2], we have now developed a method to predict the location of allosteric sites on protein structures.

The methodology, which is a coarse-grained approach based on protein flexibility, achieves up to 65% accuracy on a set of 58 different protein families. To our knowledge, this work represents the first attempt to predict allosteric sites on multiple protein families.

[1] Peracchi A, Mozzarelli A (2011) “Exploring and exploiting allostery: Models, evolution, and drug targeting.” Biochim Biophys Acta 1814: 922–933.

[2] Panjkovich and Daura, (2010) “Assessing the structural conservation of protein pockets to study functional and allosteric sites: implications for drug discovery” BMC Structural Biology, **10**:9

RNA structure prediction by knowledge-based statistical potentials and Selective 2′-Hydroxyl Acylation and Primer Extension (SHAPE).

David Dufour and Marc A. Marti-Renom

Structural Genomics Laboratory, Centro de Investigación Principe Felipe, Valencia, Spain.

New RNA structure prediction tools are needed to fast obtaining detailed structural information of new non-coding RNA sequences. Here we propose to use knowledge-based statistical potentials and low-resolution experimental evidences as input to predict RNA structure from sequence. On the one hand, we have derived a series of knowledge-based statistical potentials extracted from the X-ray RNA structures available in the PDB. Such potentials describe the C3' – C3' distances between neighboring residues and between opposed residues of a base pair, as well as the torsion angles defined between them. On the other hand, SHAPE [1] provides information about the secondary and tertiary structure of RNAs that can be translated into spatial structural restraints. Both structural data will be used as inputs in the Integrative Modeling Platform (IMP) [2] to perform simple modeling of RNA with canonical base-pairs. Moreover, we are working towards a classification of SHAPE reactivities based on the different base-pairs present in the RNA so that SHAPE can be used for refining the final RNA structure predicted by IMP.

[1] EJ. Merino *et al. Journal of American Chemical Society* **vol 127** (2005) 4223-4231.
[2] F. Alber, F. *et al. Nature*, (2007) **vol 450** 683–694.

Structure-based statistical and experimental analysis of transmembrane helices

Marc A. Marti-Renom[1], Ismael Mingarro[2] and Carlos Baeza[1,2]

[1]Structural Genomics Laboratory, Bioinformatics and Genomics Department, Centro de Investigación Príncipe Felipe, Valencia, Spain
[2]Membrane Proteins Laboratory, Department of Biochemistry and Molecular Biology, Universitat de València, Burjassot, Spain

The thickness of the hydrocarbon core of the membrane bilayer is ~30 Å, which leads to the expectation that transmembrane (TM) helices of helix-bundle membrane proteins (MPs) should be at least ~20 amino acids long to span this region. Although some studies have previously shown that 10-residue polyleucine segments can be inserted into lipid bilayers [1, 2], there are no experimental evidences on the minimum length for a TM segment to be inserted into the membrane by the translocon machinery. In the present work we have analyzed the length and amino acids composition for TM helices compared with α-helices of globular proteins to emphasize the differences between them. Using the statistical data obtained for TM helices we have designed a set of TM segments with naturally occurring amino acid distributions in TM regions of integral membrane proteins with known high resolution structure. We have then calculated the probability of insertion for the designed sequences and validated our predictions using an *in vitro* experimental system based on the *Escherichia coli* inner-membrane protein Lep as a model protein [3, 4].

[1] Chen HF and Kendall DA *J Biol Chem* **270(23)** (1995) 14115.
[2] Jaud S et al. *Proc Natl Acad Sci U S A* **106(28** (2009)**)** 11588.
[3] Hessa T et al. *Nature* **433(7024)** (2005) 377.
[4] Hessa T et al. *Nature* **450(7172)** (2007) 1026.

JORANDAS DE BIOINFORMÁTICA BARCELONA 23-25TH OF JANUARY 2012

Bacterial transcriptomic analysis using NGS

Rodrigo LOMAS, Sonia TARAZONA, Miguel Angel PARDO, Ana CONESA
Genomics of Gene Expression Lab
Centro de Investigación Príncipe Felipe de Valencia

RNASeq analysis in bacteria (1) is not only used to know new genes as the important groupe of non coding RNA (ncRNAs), it´s also used to understand the relation between a high number of genes. Our group has got the opportunity to participate in the last three years in a project of Pathomics looking for the transcriptome analysis of two diffents bacteria: Pseudomonas aeruginosa and Chamydophila pneumoniae.

After the analysis of the sequencing raw data obtained by the Biogune company using the Illumina platform and not strand specific RNASeq (ssRNASeq), we present a operonic list of the P. Aeruginosa genome, including more than 60% of all genome annotated. With this analysis we detected over 80% and also the presence of novel genes not annotated, including several putative ncRNAs.

This analysis provides the importance of NGS for the transcriptome characterization of multiple bacteria (2), including the oportunistic pathogen P. Aeruginosa. After receiving the new raw data sequenced with strand specific RNASeq (ssRNASeq) by the BGI company using the SOLID platform, the new raw data analysis is in process.

1. Science 27 November 2009: Vol. 326 no. 5957 pp. 1268-1271
2. J. Bacteriol. May 2009 vol. 191 no. 10 3203-3211

Genome-wide clustering of transcription factors by comparison of predicted protein-DNA interfaces

Alvaro Sebastian[1], Bruno Contreras-Moreira[1,2]

[1]Laboratorio de Biología Computacional, Estación Experimental de Aula Dei/CSIC,
Av. Montañana 1005, Zaragoza, España
[2]Fundación ARAID, Paseo María Agustín 36, Zaragoza, España

Transcription Factors (TFs) play a central role in gene regulation by binding to DNA target sequences, mostly in promoter regions. However, even for the best annotated genomes, only a fraction of these critical proteins have been experimentally characterized and linked to some of their target sites. The dimension of this problem increases in multicellular organisms, which tend to have large collections of TFs, sometimes with redundant roles, that result of whole-genome duplication events and lineage-specific expansions. In this work we set to study the repertoire of *Arabidopsis thaliana* TFs from the perspective of their predicted interfaces, to evaluate the degree of DNA-binding redundancy at a genome scale [1]. First, we critically compare the performance of a variety of methods that predict the interface residues of DNA-binding proteins, those responsible for specific recognition, and measure their sensitivity and specificity [2, 3]. Second, we apply the best predictors to the complete *A.thaliana* repertoire and build clusters of transcription factors with similar interfaces. Finally, we use our in-house footprintDB [4] to benchmark to what extent TFs in the same cluster specifically bind to similar DNA sites. Our results indicate that there is substantial overlap of DNA binding specificities in most TF families. This observation supports the use of interface predictions to construct reduced representation of TF sets with common DNA binding preferences.

[1] Y. Van de Peer, J. A. Fawcett, S. Proost, L. Sterck and K. Vandepoele. Trends Plant Sci **14** (2009) 680-8.

[2] B. Contreras-Moreira, P. A. Branger and J. Collado-Vides. Bioinformatics **23** (2007) 1694-6.

[3] J. Si, Z. Zhang, B. Lin, M. Schroeder and B. Huang. BMC Syst Biol **5 Suppl 1** (2011) S7.

[4] A. Sebastian and B. Contreras-Moreira. (2011)

Gene expression data in the light of elementary flux modes

Alberto Rezola[1], Adam Podhorski[1], Jon Pey[1], Angel Rubio[1] and Francisco J. Planes[1]

1 CEIT and TECNUN, University of Navarra, Manuel de Lardizabal 15, 20018 San Sebastian, Spain

The Elementary flux modes (EFMs) approach is an efficient computational tool to predict novel metabolic pathways. Elucidating the physiological relevance of EFMs in a particular cellular state is still an open challenge. Different methods have been presented to carry out this task. However, these methods typically use little experimental data, exploiting methodologies where an a priori optimization function is used to deal with the indetermination underlying metabolic networks. Available "omics" data represent an opportunity to refine current methods. In this contribution, we present a novel mathematical approach that infers a characteristic set of EFMs for a certain physiological/cellular condition based on gene expression data. Results are presented for 10 different human tissues. Its main metabolic functions are highlighted and compared. Potential applications of our approach in the light of existing methods are discussed.

GSVA: Gene Set Variation Analysis

Sonja Hänzelmann[1,2], Robert Castelo[1,2], and Justin Guinney[3]

[1]Research Program on Biomedical Informatics, Institut de Recerca Hospital del Mar, Barcelona, Catalonia, Spain
[2]Dept. of Experimental and Health Sciences, Universitat Pompeu Fabra, Barcelona, Catalonia, Spain
[3]Sage Bionetworks, Seattle, Washington, USA

Gene set enrichment (GSE) is an important type of bioinformatic analysis used to identify differentially activated gene sets or pathways in gene expression data. By providing a functionally coherent view of complex molecular systems, GSE analysis of pathways offers greater interpretability and often consistency across biological studies. Methods for GSE have focused primarily on analysis with respect to specific phenotypes, and consequently do not generalize to phenotype-independent modes of operation.

To address this limitation we introduce Gene Set Variation Analysis (GSVA), a method that transforms a gene by sample matrix to a gene set by sample matrix, enabling analyses within the 'space' of pathways instead of genes.

GSVA provides two principal advantages over other GSE methods: (1) it permits assessment of the variation of pathway activity across samples in a completely unsupervised manner, and (2) it facilitates the generation of more flexible -- and in turn, more sophisticated -- phenotypic-pathway models. We demonstrate these advantages in a series of experimental vignettes, spanning both simple techniques like clustering and survival analysis to more complex examples such as CNV-pathway and meta-pathway analyses.

Non-redundant and modular enrichment analysis for functional genomics

Daniel Tabas-Madrid[1], Ruben Nogales-Cadenas[1], and Alberto Pascual-Montano[1]

[1] *National Center for Biotechnology CNB-CSIC*

Since its first release in 2007, Genecodis [1,2] has become an indispensable tool to interpret functionally the results from experimental techniques in Genomic. Its analyses based on the *Concurrent Enrichment* of the biological terms related to lists of genes, allow not only getting functional and regulatory information connected to these genes, but also relate the information itself in a modular way. That is, GeneCodis is able to describe set of genes through modules of biological terms from different sources of information.

The numerous feedbacks of GeneCodis' users and the natural evolution of Genomics in Bioinformatics area have allowed the growth of the tool and the development of this third release. The improvements of this new version include the support of a new curated regulatory information, a compendium of miRNA-mRNA interactions from predictive and experimentally validated databases, among others new biological sources like Biocarta pathways, genetic diseases from OMIM and Genetic Association DB, protein-protein interactions data from Reactome and HPRD, even a biomedical literature analysis is allowed through the inclusion of Pubmed Ids. Now, functional exome analysis can also be done in Genecodis 3.0, in consonance with many of the recent increase of resequencing studies.

A special effort has been made to remove noisy output from the enrichment results by filtering biological terms based on the frequency of their annotations, as well as their intrinsic redundancy, inherent in this kind of data. With this goal, a new algorithm has been added to summarize the information generated by enrichment analysis sand to generate functionally coherent modules of genes and terms [3]. Additionally it is also possible to compare different results, an interesting exercise to measure in a biological way the common and divergent aspects of different experimental conditions.

Finally, it was also taken into account the graphic section of the tool, providing a user-friendly experience trough the insertion of new interactive graphics and the option to filter results when searching for specific topics of interest. Like in previous versions of GeneCodis, it's still possible to include user's data in the analyses and access all functionalities programmatically, allowing their insertion in different analysis pipelines. GeneCodis 3.0 is publicly available at http://genecodis.cnb.csic.es

[1] P. Carmona-Saez, et.al. Genome Biology. 2007 Jan 4;8(1):R3
[2] R. Nogales-Cadenas, et.al. Nucleic Acids Research 2009; doi: 10.1093/nar/gkp416
[3] C. Fontanillo, et.al. 2011 PLoS ONE 6 (9) p. e24289

Improving microRNA target prediction by performance-based algorithm combination

Ignacio Sanchez Caballero[1], Ander Muniategui[2], Ruben Nogales-Cadenas[1], Carlos O. S. Sorzano[1], Angel Rubio[2] and Alberto Pascual-Montano[1]

[1]National Center for Biotechnology-CSIC. Madrid, Spain
[2] CEIT and TECNUN, University of Navarra, San Sebastian, Spain.

MicroRNAs have been at the center stage of the biomedical community for more than a decade now, after a team of scientists working at Harvard University in Cambridge discovered their vital role in the regulation of gene expression [1]. Since then, the use of bioinformatics tools has become a major accelerator in our understanding of microRNA function. Many algorithms have been created to predict where microRNAs are encoded, as well as what genes they regulate [2]. Unfortunately, due to the popularity of the field, it is not always clear which of the available computational methods is best suited for determining which transcripts targets are regulated by which microRNAs.

We propose a straightforward way to combine the tens of currently available prediction algorithms, and assign them a credibility measure based on their previous performance to simplify the task of experimental validation. Using some additional assumptions, we have created a new database, which provides a confidence score for each predicted interaction. This score is computed taking into account the number of databases where the interaction appears, the quality of these databases in terms of their predictive accuracy, and the ranking that each database assigns to its predictions. Using cross-validation, we show that this database outperforms in terms of quantity (number of interactions) and quality (ability to predict experimentally validated interactions) any of the previous ones.

No algorithm makes perfect predictions under every condition. Because of the multi-faceted nature of miRNA targeting, and the lack of consensus among existing predictions, it makes sense to combine them in a way that maximizes the number of validated predicted results. There have been previous attempts to combine the predictions of several algorithms by first taking their union or intersection as a way to improve coverage or accuracy respectively, balancing out their sensitivity and specificity, and then choosing the most likely candidates by consensus [3]. Most of these algorithms give the user the ability to choose which combination of databases should be used. The problem with this approach is that in a significant proportion of cases we do not have the necessary information about each database's performance to make an informed decision. Our approach presents an alternative solution that assigns confidence scores to each database's predictions. This solves the problem introduced by choosing candidates by consensus; mainly, that several low-confidence predictions for the same interaction can erroneously appear as more credible than a single high-confidence prediction.

A new database containing a compendium of scored miRNA-mRNA interactions will also be presented.

[1] Lee, R.C., et. al. *Cell* **75** (1993), pp. 843-854
[2] Gaidatzis, D., et. al. *BMC Bioinformatics* 8 (2007), p. 69.
[3] Megraw, M. *Nucleic Acids Research*, 35 (2007), pp. 149-155

ANNOTATION PIPELINE FOR 454 ASSEMBLY OF SEA BASS GONADS USING MODEL FISHES AND ANOTHER DATABASES

D. Xavier[1*], JM Rodríguez[2], F. Moran[1], B. Crespo[3], A. Gómez[3], L. Ribas[4], F. Piferrer[4]

[1] Departamento de Bioquímica y Biología Molecular I, Universidad Complutense Madrid, Avd. Complutense s/n, 28040 Madrid, Spain.
[2] Spanish National Bioinformatics Institute (INB), INB Central Node, 28029 Madrid, Spain.
[3] Fisiología de la Reproducción de Peces, IATS, CSIC. 12595 Castellón, Spain.
[4] Biología de la Reproducción, ICM, CSIC. 08003 Barcelona, Spain.
[*] Email: **danidiasxavier@quim.ucm.es**

Abstracts

Sea bass (*Dicentrarchus* labrax) is one of the most relevant fish for marine aquaculture in Spain and has been extensively studied by Aquagenomics Project (http://www.aquagenomics.es). In this work, we present the annotation and curation pipeline developed to functionally annotate sea bass transcriptome. The genetic data was obtained from two cDNA libraries of gonads at different stages of maturation by experimental groups of Aquagenomics Project (IATS and ICM, CSIC, Spain). The 454 sequencing was performed in the Max Planck Institute for Molecular Genetics and the obtained reads were assembled with MIRA [1], yielding a total of 32,427 contigs.

The annotation was carried out with respect to model fishes from Ensembl [2] database (*Danio rerio*, *Tetraodon nigroviridis*, *Gasterosteus aculeatus*, *Oryzias latipes, Takifugu rubripes*) and NCBI [3] no-redundant protein database. For both databases, we performed BLAST [4] searches to distinguish the different threshold levels of gene/protein homology. Besides, the pipeline executes PfamScan [5] algorithm and RPS-BLAST against NCBI Conserved Domains Database [6] to detect possible protein domains. Finally, genes were mapped to Gene Ontology (GO) terms [7] using cross-reference data from Gene Ontology, NCBI, Ensembl and UniProt [8] databases, and also through BLAST2GO [9] software.

The curation process was developed based on rules that give different scores to each decision and take into account aspects such as the existence of domains, the conservation of the active sites, e-value, identity percentage, and homology length.

The results from annotation and curation will be presented and discussed in terms of their application to DNA-microarray design in the framework of the Aquagenomics Project.

Acknowledgements

This work has been supported by Grant CSD2007-00002 Consolider-Ingenio 2010, MICINN (Spain)

[1] Using the miraEST Assembler for Reliable and Automated mRNA Transcript Assembly and SNP Detection in Sequenced ESTs. *Genome Res.* Jun.14 (6):1147-59, 2004
[2] The Ensembl genome database project. *Nucleic Acids Research* . 30(1):38-41, 2002
[3] National Center of Biotechnology, http://www.ncbi.nlm.nih.gov/.
[4] Basic local alignment search tool. *J Mol Biol*. Oct 5;215(3):403-10, 1990
[5] The Pfam protein families database. *Nucleic Acids Research*. Database Issue 38:D211-222, 2010
[6] CDD: a Conserved Domain Database for the functional annotation of proteins. *Nucleic Acids Res*. Doi:10.1093, 1-5, 2010
[7] Gene Ontology: tool for the unification of biology. *Nature Genetics* 25, 25-29, 2000
[8] The UniProt Consortium. Ongoing and future developments at the Universal Protein Resource. *Nucleic Acids Res*. 39: D214-D219 (2011)
[9] Blast2GO: A universal tool for annotation, visualization and analysis in functional genomics research. *Bioinformatics* 21: 3674-3676, 2005

Gene Functional Annotation beyond Enrichment Analysis: moving from gene lists to gene annotated metagroups and gene networks

Celia Fontanillo[1], Sara Aibar[1], Ruben Nogales-Cadenas[2], Alberto Pascual-Montano[2]and Javier De Las Rivas[1]

[1]Bioinformatics and Functional Genomics, Centro de Investigación del cáncer (CIC-IBMCC, CSIC/USAL) Salamanca, Spain
[2]National Center of Biotechnology (CNB, CSIC), Campus de Cantoblanco UAM, Madrid, Spain

Functional analysis of large sets of genes and proteins is becoming more and more necessary due to the increase of experimental biomolecular data at omic-scale. Enrichment analysis (EA) is by far the most popular available methodology to derive functional implications of sets of cooperating genes. This type of analysis searches in biological databases of gene attributes –such as Gene Ontology (GO) or KEGG pathways– and uses statistical testing to find significant annotations assigned to specific genes. However these kind of methods do not address several issues related to the gene annotations, such us: redundancy of biological terms repeated in many databases (e.g. cell cycle GO:0007049, cell cycle KEGG hsa04110, etc), bias towards highly-frequent unspecific terms (e.g. GO:0050789 "regulation of biological process" includes more than 44% of all human genes annotated), or inadequate functional annotation of genes (e.g. NRAS is not annotated to GO:0043410 "positive regulation of MAPKKK cascade").

To overcome these limitations we have developed a computational method that filters the output of an Enrichment Analysis and finds significant and coherent groups of genes and terms. With these metagroups we are able to summarize and facilitate the interpretation of the biological functions and processes represented in an initial query gene list, and also inferring new associations among genes that may be cooperating in a process which has not been annotated yet. The association of each gene to these "functional metagroups" is used, in a further step, to calculate a "functional score" for each gene. Based on the comparison of these scores we can estimate a "functional distance" for each gene-pair in the initial gene list. Using this method we are able to build gene networks that reflect the proximity of each gene-pair based on a comprehensive analysis of their biological annotations.

The method has been tested with a small set of well-known interacting proteins from yeast and with a large collection of reference sets from three heterogeneous resources: mamalian multi-protein complexes (CORUM), yeast cellular pathways (SGD) and human diseases (OMIM). Moreover, we have also tested the algorithm with the gene signatures of 4 classes of leukemia obtained by analysis of their expression profiles. Our tool, named "*GeneTerm Linker*" (http://gtlinker.cnb.csic.es) produces robust results even introducing different levels of noise in the genes used for the tests. The potential to provide gene networks based on biological functional annotations is also demonstrated in this communication.

[1] Fontanillo C, Nogales-Cadenas R, Pascual-Montano A, De las Rivas J. (2011). Functional analysis beyond enrichment: non-redundant reciprocal linkage of genes and biological terms. *PLoS One*, 6(9), e24289.

An integrative approach to the detection of cancer driver genes

Abel Gonzalez-Perez[1] and Nuria Lopez-Bigas[1,2]

[1]*Research Group on Biomedical Informatics-GRIB, Universitat Pompeu Fabra, Barcelona Biomedical Research Park-PRBB, Barcelona, Spain*
[2]*Institució Catalana de Recerca i Estudis Avançats (ICREA), Barcelona, Spain*

Cancer genome sequencing projects [1,2] are uncovering large lists of alterations across cohorts of samples of several tumor types. Although these efforts have opened up the possibility to identify which specific genomic and epigenomic alterations are essential for the development of each tumor type, this issue is far from resolved. Some approaches have been taken in the past couple of years to uncover such cancer *driver* alterations and the genes that bear them in a few tumor types. Typically, such approaches rely either on a) an evaluation of the potential impact of genomic alterations on protein expression or function [3,4] or b) an assessment of the statistical over-representation of a given gene within the alterations list of a tumor genome [5,6]. Recently, this second type of –recurrence based– strategies have been used in conjunction with other expected features of driver genes, such as functional closeness and mutual exclusivity [7,8].

Nevertheless, these recurrence based approaches normally fail to take into account differences in background alterations rate between samples. Moreover they are unable to detect lowly recurrently altered genes which may be important to tumorigenesis. On the other hand, functional impact approaches give no information whatsoever on the frequency of alterations and therefore are prone to underestimate marginally deleterious mutations with a potential for driverness.

We have approached the problem of detecting potential driver genes from large cohorts of tumor genome samples by combining four different sources of information. First, we evaluate the functional impact of non-synonymous SNVs (nsSNVs) on genes by combining three predictive methods using a variation of an idea recently developed in our group [9] and identify those that tend to accumulate functional nsSNVs. Second, we identify which genes are targets of nsSNVs and/or Copy Number Variations (CNVs) more frequently than expected by chance. We also use the enrichement for functional connections between likely driver genes (in the previous list) and knowledge available from at least one previous independent study. We demonstrate our integrative approach using the catalog of genomic alterations identified in cohorts of samples of glioblastoma multiforme and serous ovarian carcinoma.

[1] The Cancer Genome Atlas. *Nature* **455** (2008).
[2] International network of cancer genome projects. *Nature* **464** (2010) 993-998.
[3] Reva B et al. *Nucleic Acids Research* (2011).
[4] Carter H et al. *Cancer Research* **69** (2009) 6660-6667.
[5] Greenman C et al. *Nature* **446** (2007) 153-158.
[6] Akavia UD et al. *Cell* **143** (2010) 1005-1017
[7] Cerami E et al. *PLoS ONE* **5** (2010).
[8] Ciriello G et al. *Genome Research* (2011).
[9] Gonzalez-Perez A et al. *AJHG* **88** (2011) 440-449.

COVER: *A priori* estimation of coverage for metagenomic sequencing

Javier Tamames, Santiago de la Peña and Victor de Lorenzo

Centro Nacional de Biotecnología (CNB-CSIC). C/Darwin 3, 28049 Madrid (Spain)

In any metagenomic project, the coverage obtained for each particular species depends on its abundance. This makes it difficult to determine *a priori* the amount of DNA sequencing necessary to obtain a high coverage for the dominant genomes in an environment. To aid the design of metagenomic sequencing projects, we have developed COVER, a web-based tool that allows the estimation of the coverage achieved for each species in an environmental sample. COVER uses a set of 16S rRNA sequences to produce an estimate of the number of OTUs in the sample, provides a taxonomic assignment for them, estimates their genome sizes and, most critically, corrects for the number of unobserved OTUs. COVER then calculates the amount of sequencing needed to achieve a given goal. Our tests and simulations indicate that the results obtained through COVER are in very good agreement with the experimental results.

Non-Ribosomal Peptide Synthase Substrate Predictor

Prieto C[1], García-Estrada C[1], Lorenzana D[1] and Martín JF[1]

[1] *Instituto de Biotecnología de León (INBIOTEC), Parque Científico de León, Av. Real, 1, 24006 León, Spain.*

NRPS enzymes are multi-modular enzymes involved in the biosynthesis of natural products. A minimal non-ribosomal peptide synthetase module contains specific functional domains which are able to catalyse several activities, such as amino acid adenylation (A- activation domain, determining the substrate specificity through a code of amino acid residues found inside the adenylation domain itself), thioesterification (T thiolation or acyl carrier domain) and peptide-bond formation (C- condensation domain), allowing elongation of the nascent peptide (Schwarzer and Marahiel 2001).

Not many bioinformatics studies have been done in this area. Early studies established critical residues in NRPS A domains and general rules for deducing substrate specificity. These investigations used the crystal structure of the peptide synthetase GrsA, which was solved in binding a phenylalanine, in order to define a specificity conferring code of residues in the A domain based on the distance between them and the binding substrate. However, these methods have problems in classifying fungal NRPSs, because the GrsA crystal seems to be an inadequate model for them.

In connection to this, Khurana *et al.* (2010) have applied HMM to functionally classify the acyl:CoA synthetase super-family members. This work suggests that the application of HMM profiles to classify this superfamily, outperforms the predictions based on the limited number of active site residues (Khurana *et al.* 2010). This conclusion can also be applied to a more ambitious goal, such as the determination of the substrate which binds to an adenylation domain.

The current *omics* era has enabled the exponential growth of the sequenced NRPS. This implies that a tool which could predict the specificity of their A domains is of increasing interest, and its training could be beneficial with the new annotated NRPS. Based on these facts, the previous experience of our group in the area and the cited publications, we have launched the present work, whose ultimate goal is to develop a new bioinformatic tool to achieve the prediction of substrates which bind to adenylation domains in NRPS. The approach that we present improves the performance and reliability of previous ones, overcomes the problem with fungi proteins, and has been implemented in a useful software tool for the functional analysis of incoming NGS data.

Availability: The database and the predictor are freely available on an easy-to-use website at www.nrpssp.com.

[1] Schwarzer D et al. *Naturwissenschaften* **88** (2001) 93-101.
[2] Khurana P et al. *BMC Bioinformatics* **11** (2010) 57.
[1] Prieto C et al. *Bioinformatics* (2011) doi: 10.1093/bioinformatics/btr659.

Using high-throughput technologies to improve stress tolerance characteristics in sunflower

Francisco García-García[1,2], Paula Fernández[3], Ana Conesa[1], Norma Paniego[3], Ruth Heinz[3], Lucila Peluffo[4], Veronica Lia[4], Laura de la Canal[5], Karina F. Ribichich[6], Raquel Chan[6], Julio Di Rienzo[7] and Joaquín Dopazo[1,2]

[1]Functional Genomics Node, National Institute of Bioinformatics, CIPF, Valencia, España
[2]Department of Bioinformatics, Centro de Investigación Príncipe Felipe, Valencia, España
[3]Instituto de Biotecnología, CICVyA-INTA Castelar, Buenos Aires, Argentina
[4]Comisión Nacional de Investigaciones Científicas y Técnicas – CONICET, Argentina
[5]Universidad Nacional Mar del Plata, Argentina
[6]Universidad Nacional del Litoral, Santa Fe, Argentina
[7]Universidad Nacional de Córdoba, Córdoba, Argentina

Sunflower (*Helianthus annuus L.*) is one of the most relevant crops as a source of edible oil. Advances in sunflower genomics have greatly enhanced the development and application of new tools for crop improvement, and promoted the expansion of sunflower uses to new markets like biofuels, biolubricants and biopharma. In this context, the power of throughput technologies allow us to bring new insights into the genomic information that would become a key tool to afford an efficient system for molecular breeding.

Gene expression chip was designed specifically for sunflower. After its validation, microarray analysis has been performed concerning various stress situations: water deficit as a physiological event that induces senescence, resistance to sclerotinia, application of root-modifying treatments (jasmonic acid, ibuprofen) and application of endogenous transcription factors. Blast2GO [1] has been used to generate the functional annotation and the suite of tools, Babelomics [2], allowed us to analyze microarray data.
RNA-seq assays will include a study of quantification of expression (Illumina) and an observational study with standard libraries (454) to build a catalog of transcripts which will be used to assemble a reference genome for the mapping. The first study will be on different levels of resistance to sclerotinia.

This work generated the first custom sunflower oligonucleotide-based microarray under Agilent technology, validated for its application to transcriptional studies of sunflower subjected to different growth and/or developmental conditions.

The use of high-throughput technologies allow us functional genomic characterization of sunflower. This information is useful to improve the characteristics of tolerance to biotic and abiotic stress.

[1] Babelomics: an integrative platform for the analysis of transcriptomics, proteomics and genomic data with advanced functional profiling. Ignacio Medina et al. NAR-00461-Web-B (2010).

[2] Blast2GO: a universal tool for annotation, visualization and analysis in functional genomics research. Ana Conesa et al. *Bioinformatics*, **21**, pp. 3674-3676 (2005).

PEANUT: A comprehensive tool for protein expression analysis

Rodríguez-Moreno A[1], Cano-Castillo M[1], Goikoetxea MJ[2], Sanz ML[2]and Trelles O[1]

[1]Computer Architecture Department, University of Malaga, Malaga, Spain
[2]Department of Allergology and Clinical Immunology, University Clinic of Navarra, Pamplona, Spain

Background: The increase in the availability of recombinant allergens together with the appearance of protein microarrays has made of allergen arrays a very powerful tool in allergy diagnosis [1]. Allergen arrays consist of a solid surface with a range of spotted allergens in which a small amount of serum is added. Thus, the detection of specific IgE against a particular allergenic protein results in a more precise diagnosis of sensitization and this goal can be achieved by using this type of microarrays. The mentioned technique comercially available, called ImmunoCAP ISAC®, is one of those that allow a component-resolved diagnosis. This term refers to the recognition of a particular component contained in an allergenic source which is the real trigger of an allergic reaction [2].

As for gene microarrays, it is important to bear in mind the many sources of systematic and random errors introduced in these techniques that could perturb our capabilities to detect the authentic biological variation [3].

The majority of the hardware and software for the analysis of protein arrays has been adapted from DNA microarray tools [4]. Although this has been useful in the early stage of protein arrays, more tailored tools are needed to deal with the specific issues [5]. In this paper we present a user-friendly software focused in the analysis of ISAC®arrays. Apart from having added quality filtering methods the software here offered also contains a 2Scan procedure aimed at extending the dynamic range of fluorescent intensities, thus avoiding saturation and quantization problems, strategy that has yielded good results before [6,7,8].

Results: PEANUT is a standalone application with a friendly interface which integrates a complete set of tools for allergen arrays analysis, ranging from computer vision techniques for spot segmentation and quantification, to graphical representations for data analysis and interpretation. It allows spot filtration, replicate resolution and error removal, including All-in-one-Click procedures for fast and automated executions. Additionally, a unique implementation of the double scan procedure is available. PEANUT automatically detects slides captured at low and high intensities and combines both measurements to produce a high dynamic range (HDR) measurement extending the hardware scanner capabilities and signal quality. The application allows customizing a wide set of parameters for adjusting to the user requirements. It is possible to add new extensions to include new protocols and formats and to adapt to different data acquisition devices when necessary.

Acknowledgments: This work has been partially supported by Grant RD07/0064/0017 RIRAAF - from the ISCIII-Spain. MC and OT partially work for the GN5 at the National Institute for Bioinformatics, Spain

References

[1] Schreffler WJ. Journal of Allergy and Clinical Immunology **127** (2011): 843-849.
[2] Goikoetxea et al. Allergologia et Immunopatologia [Madrid] **38** (2010): 37-40.
[3] García de la Nava et al. Statistical applications in genetics and molecular biology **3** (2004): Issue 1, Article11
[4] Lubomirsky et al. Journal of Computational Biology **14** (2007): 350-359.
[5] Sboner et al. Journal of Proteome Research **8** (2009): 5451-5464.
[6] Deinhoferet al. Methods **32** (2003): 249-254.
[7] Jahn-Schmid et al. Clinical and Experimental Allergy **33** (2003): 1443-1449.
[8] Wöhrl et al. Allergy **61** (2006): 633-639.

Biochemical features associated to cancer mutations

E. Porta[1], AM. Rojas[1], and I. Cases[1]

[1]*Computational Cell Biology Group, Institute for Predictive and Personalized Medicine of Cancer, Badalona, Spain*

The development and popularization of genome-wide technologies to perform experiments has provided the scientific community tools capable of taking a broad picture of the events happening in a cell. In this scenario, computational approaches that rely on computer-friendly annotations, such as controlled vocabularies or ontologies, to analyze the data have proven useful. The most widely used ontology to perform these studies is the Gene Ontology (GO). This approach has been particularly useful in the case of complex phenotypes, where focusing on single genes can be misleading since there are many genes involved. In these cases, analyzing the data from a broader perspective by aggregating the information using the ontology structure can be useful in order to extract new insights in the biology of the disease and learn about the differences in the genes underlying different phenotypes.

One particular disease that has benefited from these approaches is cancer, where several GO terms (such as “Apoptosis”, “DNA repair”, “Intracellular signaling” etc) have been identified to be over or under-represented in its associated genes in different enrichment analysis[1]. These associations have been used to make predictions on new cancer-related genes or to infer new functional annotations in known cancer-associated genes [1].

Enrichment analysis is not restricted to the Gene Ontology, however there have been few attempts to use this approach with other biological ontologies. We think that given the increasing interest in next-generation sequencing techniques and the exponential growth in the number of genomes and mutations coming from cancer samples it would be interesting to extend the enrichment analysis to this type of data because, not only different genes cause different diseases, also different mutations in the same gene can be associated to different phenotypes [2].

We have used the Sequence Ontology[3] (SO) and the Disease Ontology (DO) to perform an enrichment analysis in a dataset of human disease-associated missense mutations coming from Online Mendelian Inheritance in Man (OMIM), the Genetic Association Database (GAD) and the Catalogue Of Somatic Mutations In Cancer (COSMIC) to identify under or over-represented SO terms in cancer-related mutations.

Using this approach we have identified 3 SO terms enriched in cancer-associated mutations (“serine-rich regions”, “aminoacid-biased regions” and “intrinsically unstructured regions”) and 3 other SO terms depleted in cancer-related mutations (“disulphide bridges”, “peptide localization signals” and “transmembrane regions”) that highlight different relevant features of the disease.

1. Hu, P., Bader, G., Wigle, D. a & Emili, A. Computational prediction of cancer-gene function. *Nature reviews. Cancer* **7**, 23-34 (2007).
2. Zhong, Q. et al. Edgetic perturbation models of human inherited disorders. *Molecular systems biology* **5**, 321 (2009).
3. Eilbeck, K. et al. The Sequence Ontology: a tool for the unification of genome annotations. *Genome biology* **6**, R44 (2005).

A method to asses the common core of a set of proteomes

Oscar Conchillo-Solé[1], Daniel Yero[1], Isidre Gibert[1], and Xavier Daura[1,2]

[1]*Institut de Biotecnologia i de Biomedicina "Vicent Villar Palasí" (UAB)*
[2]*ICREA*

The development of efficient and inexpensive genome sequencing methods has produced lots of complete proteomes deposited in multiple repositories, some of them belonging to different strains of the same species. This forces scientists studding one or some of these organisms' proteins to take into account not just one, but all the equivalent ones in the other strains [1]. Saying that entails that there is a pool of proteins common in all the strains of a species; this set is generally known as core proteome [2]. Here we present a method (and a tool) that for a given set of proteomes easily finds these proteins and lists them together with their closest ones for the other proteomes. Our method relays in the capability of comparing proteomes of the UCSC blat [3] program and uses a complete graph [4] approach of all their best matches to accept or reject groups of proteins to the core proteome.

[1] Hervé Tettelin et al. PNAS **102** (2005) 13950.
[2] Stephen J. Callister et al. *PLoS ONE* **3** (2008) e1542.
[3] Kent WJ. *Genome Res.* **12** (2002) 656.
[4] Narsingh Deo *Graph Theory with Applications to Engineering and Computer Science* **chap. 2** (2004) 32

The impact of small versus large chromosomal rearrangements in sequence and expression divergence

Rui Faria[1,2], Jordi Rambla[1], Sandra Neto[2], Katja Nowick[3] and Arcadi Navarro[1,4]

[1]IBE, Institute of Evolutionary Biology (UPF-CSIC), Departament de Ciències de la Salut i de la Vida. Universitat Pompeu Fabra. PRBB, Barcelona, Spain
[2]CIBIO, Research Center in Biodiversity Genetic Resources, Universidade do Porto. Campus Agrário de Vairão, Vairão, Portugal
[3]Bioinformatics Department, Universität Leipzig, Leipzig, Germany
[4]Institució Catalana de Recerca i Estudis Avançats (ICREA) and Universitat Pompeu Fabra, Barcelona, Spain

Suppressed-recombination models of chromosomal speciation suggest that chromosomal rearrangements (CRs) can act as genetic barriers by reducing recombination. This provides an opportunity for the accumulation of genetic incompatibilities and consequently, the establishment of reproductive isolation between populations despite the occurrence of gene flow. A testable prediction of these models is that sequence and gene expression divergence are expected to be higher within CRs, when compared with collinear regions. Different outcomes concerning the importance of rearrangements' size in speciation can be hypothesized: large CRs can bear a higher number of genes (or larger coding sequence) increasing the chance of accumulating incompatibilities, but may be less effective barriers to gene flow (double cross-over and gene conversion); by the opposite, small rearrangements have a lower probability of carrying genes involved in incompatibilities, but may impose stronger recombination suppression (pairing may be more difficult). Therefore, the effect of each type of CRs in speciation cannot be easily predicted and needs to be properly tested.

Most empirical studies trying to evaluate the role of CRs in speciation were based on large CRs, reflecting an experimental bias for detecting larger fragments of the genome that rearranged. However, recent studies have highlighted the putative importance of small CRs and single gene movement on evolution, particularly on speciation. In this study, we attempt to address the contribution of large versus small CRs in speciation using two complementary approaches: empirical data and simulations. For the latter, a speciation simulator was built to test the effect of inversions' size on the maintenance of differentiation between two previously isolated populations, after a secondary contact. This approach was complemented by a comparative genomic analysis through the: i) identification of large and small CRs in several primate species (human, chimpanzee, orang-utan and macaque) based on the Ensembl gene coordinates and validated with experimental data; ii) assessment of sequence and expression divergence for each lineage; and iii) evaluation of the genomic distribution of divergence rates, especially comparing within versus outside CRs, in order to test the hypotheses mentioned above. Results and their evolutionary implications will be discussed within the framework of chromosomal speciation models.

Exome characterization of gene variants related to the fertility status of different human populations

Patricia Díaz-Gimeno[1], José Carbonell[1], Jorge Jiménez[1] and Joaquín Dopazo[1].

[1]*Bioinformatics and Genomics Department. Centro de Investigación Principe Felipe (CIPF), Valencia, Spain.*

Genome research has been revolutionized by Next-Generation Sequencing (NGS) technologies. From this perspective, Biomedicine is increasingly using them, and in particular the exome analysis, to associate genetic variants with phenotypes. Endometrial receptivity phenotype is the stage in the menstrual cycle when the embryo implantation may occur. Research regarding human reproductive health is a matter of interest in human society, however it has not yet been studied deeply using these revolutionary approaches.

The main objective of this study was to characterize the exome variants of endometrial receptivity phenotype genes in control human populations. We analyzed the exome of samples from 14 populations of the latest *1000 Genomes* release [1] and the Andalusian population from the Medical Genome Project [2]. In order to define the normal endometrial receptivity variant profile, each population has been characterized and compared within the others.

These genes were characterized and analyzed using biological databases and Babelomics platform [3]. A functional enrichment and a network gene analysis were performed to study the gene biological system. Finally, we have investigated this fertility status under a Systems Biology perspective, establishing the relationship between genetic variants and topological parameters of this endometrial receptivity network.

[1] The 1000 Genomes Project Consortium. *Nature* **467** (2010) 1061-73.
[2] Medical Genome Project http://www.medicalgenomeproject.com/
[3] Medina et al. *Nucleic Acids Res.*(2010) W210-3.

Is gene duplication rerunning the tape?

M. Mar Albà[1], and Cinta Pegueroles[1]

[1]*Evolutionary Genomics Group, Research Programme on Biomedical Informatics, FIMIM-UPF, PRBB, Barcelona, Spain.*

Genome wide studies have demonstrated that gene duplication is a major mechanism modeling genomes, specially those of vertebrates and plants. Despite their contribution to genome evolution and adaptation is widely accepted since early 70's [1], the study of their fate after their origin is still ongoing. Once the duplication has raised, the duplicated gene can be pseudogenized through the accumulation of deleterious mutations, or it can be preserved according to three main mechanisms: neofunctionalization (gain of a new function), subfunctionalization (split of the ancestral function) and increased gene dosage advantage. According to Innan and Kondrashov [2] , the three evolutionary models can be distinguished by differences in the ratio of the number of nonsynonymous substitutions per nonsynonymous sites and the number of synonymous substitutions per synonymous sites (dN/dS) between pre and postduplicated copies.

The main goal of our study is to determine whether the model in which a gene duplicate falls is mainly determined by intrinsic characteristics, such as gene sequence and function, or whether it strongly depends on the mutations accumulated at the first stages of the gene duplication. In the first case we expect a significant higher frequency of two independent duplicates of the same gene evolving under the same model than expected by chance, contrary to the second case. To try to elucidate this issue, we have focused on three data sets. The test set allows to estimate the frequency of two independent duplicates of a certain gene in *Homo sapiens* and *Mus musculus* evolving under the same model. The two control sets are used to characterize the frequency of the different models in the *H. sapiens* and the *M. musculus* genomes respectively. By comparing the frequencies of the three sets we will determine whether patterns observed in the test set could be explained by chance. We have developed a pipeline to classify duplicates into the three models, based on the comparison of the dN/dS values in the ancestral and duplicated copies by using a Fisher exact test. However, due to the low number of trees with the desired topology obtained, it seems necessary to increase our data set in order to perform more robust predictions.

[1] S. Ohno *Evolution by gene duplication* (1970).

[2] H. Innan & F. Kondrashov *Nature Reviews Genetics* **11** (2010) 97-108.

Evidence for the uniform model of evolution

Arnau Mir[1], Francesc Rosselló[1], and Lucía Rotger[1]

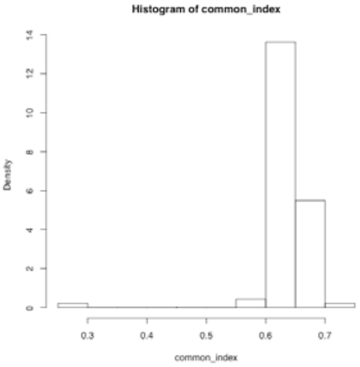
[1]Dept. Mathematics and Computer Science, Univ. of the Balearic Islands, 07122 Palma (Spain)

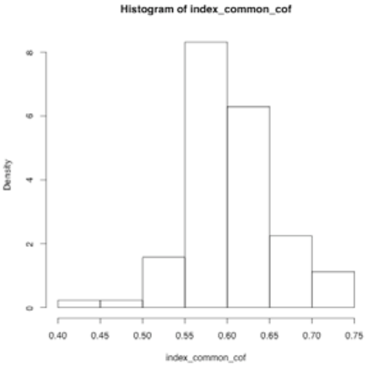
An interesting topic of research in phylogenetics is the study of what inferences about evolution can be drawn from the shape of real phylogenetic trees. For instance, recent papers have shown that phylogenetic trees are more unbalanced than the Yule model predicts [1], the degree of this imbalance having been tested on TreeBase [2]. In this paper we test on TreeBase the uniform model of evolution using nodal distances and cophenetic values.

For every pair of different leaves a,b in a rooted phylogenetic tree T, let $d_T(a,b)$ denote the *nodal distance* between a and b (the sum of the lengths of the paths from their least common ancestor to a and b), and $\varphi_T(a,b)$ the cophenetic value of a and b (the depth of their least common ancestor). Consider the random variables d_n and φ_n that give the value of $d_T(a,b)$ or $\varphi_T(a,b)$, respectively, for an equiprobably randomly chosen pair of leaves a,b on a phylogenetic tree with n leaves randomly chosen under the uniform distribution (i.e., assuming that all trees are equally likely). We have recently obtained explicit formulas for the distribution of d_n [3] and φ_n [4].

Let $\mathrm{med}(d_n)$ and $\mathrm{med}(\varphi_n)$ denote their medians. For every pair of leaves a,b of a tree T with n leaves, let $ic_d(a,b,T)=2F_{d,n}(d_T(a,b))$ if $d_T(a,b)\leq \mathrm{med}(d_n)$, and $ic_d(a,b,T)=2(1-F_{d,n}(d_T(a,b)))$ if $d_T(a,b)>\mathrm{med}(d_n)$, and define $ic_\varphi(a,b,T)$ in a similar way, replacing everywhere d by φ. If the distribution of d (resp. φ) on T ressembles d_n (resp. φ_n), then the values of $ic_d(a,b,T)$ (resp. $ic_\varphi(a,b,T)$) will be greater than 0.5, while, otherwise, they will be much smaller. So, the average of $ic_d(a,b,T)$ or $ic_\varphi(a,b,T)$, for a large number of pairs of leaves *(a,b)*, will give a hint about how close is T to a ``common'' tree from the point of view of the distribution of d_n or φ_n.

We have extracted a sample of 2000 resolved trees from TreeBase, with a range of numbers of leaves n between 3 and 91. For each tree T, we have randomly generated 100 pairs of leaves (a,b), we have computed the corresponding indices $ic_d(a,b,T)$ and $ic_\varphi(a,b,T)$ and, finally, the mean $ic_d(T)$ and $ic_\varphi(T)$ of the corresponding indices for all pairs of leaves. It turns out that most of these values lie in or close to the interval (0.55,0.65), as the histograms below show. This entails that most trees in our study have a distribution of nodal distances and cophenetic values similar to the theoretical ones under the uniform distribution. We have also performed other experiments that show no clear dependency of these indices on n.





[1] D. Aldous. *Statist. Sci.* **16** (2001), 23-24
[2] M. Blum, O François. *Syst. Biol*. **55** (2007), 685-691
[3] A.Mir, F. Rosselló. *J. Math. Anal. Appl*. **371** (2010), 168-176
[4] A.Mir, F. Rosselló, L. Rotger. In preparation.

The phylogenomics of carbohydrate metabolism in Protozoa

Joana A. Lima de Oliveira[1], Diogo A. Tschoeke[1], Salvador Capella-Gutierrez[2], Alexandros Pittis[2], Marina Marcet-Houben[2], Toni Gabaldón[2], Alberto M. R. Dávila[1]
[1]*Computational and Systems Biology Pole, Oswaldo Cruz Institute, FIOCRUZ*
[2]*Center for Genomic Regulation (CRG) and UPF, Barcelona, Spain*

Protozoa are a phylogenetically-diverse assemblage of unicellular eukaryotes that spans different eukayotic domains. Among the more than 200,000 described species of protozoa, around 10,000 display a parasitic lifestyle. The evolution of carbohydrate metabolism (CM) in Protozoa has received little attention as compared to that of model organisms. Nevertheless, previous studies have shown evidence for the existence of horizontal gene transfer (HGT) of CM genes in some protozoans, for instance glycolictic enzymes in parasitic trypanosomes. CM in Protozoa is related to energy production and development, which in turn are important processes for pathogenicity and obvious candidates for new drug targets. In this study, the proteome of 22 protozoan species obtained from NCBI and ProtozoaDB were compared to establish orthology and paralogy relationships in genes related to carbohydrate transport and metabolism ("G" functional category, according to COG/KOG-NCBI), with the aim of generating new information on the phylogenetic and evolutionary relationships of those genes. After redundancy removal, using cd-hit, of the proteomes from the genera *Plasmodium, Entamoeba, Trypanosoma, Leishmania, Giardia, Theileria, Toxoplasma, Trichomonas* and *Cryptosporidium* a total of 204,624 proteins were submitted to the OrthoMCL program for homology inference generating 26,101 homologs groups (4,982 paralogs and 21,119 orthologs). After that, all COG/KOG clusters that belongs to CM were downloaded from NCBI, in a FASTA format, and used for the functional categorization of each of the Protozoa orthologs identified. Briefly, a similarity analysis using Blast and RpsBlast was performed with the Protozoa orthologs against the CM of COG/KOG database. Using COG we identified 1,921 orthologous groups related to Prokaryotes and 1,834 orthologous groups were identified with KOG related to Eukaryotes, all of them related to G category. The best hit (higher score and lower e-value) of Protozoa to each COG and KOG was identified, filtered (a blast e-value cutoff of 1e-5 was applied) and grouped as follows (a) best blast hits to Prokaryotes [COG] = 237 orthologs and (b) best blast hits to Eukaryotes [KOG] = 209 orthologs. Best hits to Prokaryotes (COG) revealed that 5,5% (13/237) of the orthologs are more related to *Nostoc sp.*, an MC-producing organism, part of Cyanobacteria phylum which has been recently reported to be abundant in extreme environments. Best hits to Eukaryotes (KOG) revealed that 39% (82/209) of the orthologs are more similar to *Arabidopsis thaliana*. Furthermore, we made an intersection between the number of total sequences obtained with COG using OrthoMCL and the number of total sequences obtained with KOG with the intention of have two sets mutually exclusive and those OrthoMCL sequences that appears in both cases. This intersection revealed that the number of hits obtained with KOGs were equal to *Arabidopsis thaliana* and *Homo sapiens* species producing 34,48% of the matches. The hits to COG generated a huge sort of information and indicated that *Synechocystis sp.* have 11.67% of identity with the ortholog genes selected for the “G” category. The intersection of best hits to COG and KOG shows that genes such as dihydroxyacetone kinase 1, NAD(p)-dependent steroid dehydrogenase-like protein, glucose-6-phosphate 1-dehydrogenase and triosephosphate isomerase are shared by Protozoa and model organisms available at COG/KOG. Gene phylogenies were inferred using the Protozoan orthologs and paralogs with the PhylomeDB pipeline [2], and compared to species tree [1]. This trees comparison, carried out using treeKO [3], showed that 44.83% (78/174) genes of the Protozoa CP display the same topology, independently the number of paralogous/orthologous sequences present in the single trees, than the species tree, among them: pteridine transporter, alpha glucosidase, glucose transporter, nucleoside diphosphatase, phosphoglucomutase, ribokinase, inositol monophosphatase, aldose-1-epimerase, galactokinase, etc.

[1] Ocaña and Dávila et al. Evol Bioinform **7** (2011) 107-21
[2] Huerta-Cepas,J. et al. Nucleic acids research, 39 (2011), D556-60.
[3] Marcet-Houben,M. and Gabaldón,T. Nu [illegible] ids research, 39 (2011), e66.

How often does natural selection targets multiple, interacting genes? The prevalence of epistasis in recent human evolution

Natalia Petit[1] and Arcadi Navarro[1,2,3,4]

[1] Institut de Biologia Evolutiva (CSIC-UPF), [2] Departament de Ciències Experimentals i de la Salut (DCEXS). Universitat Pompeu Fabra. [3] National Institute for Bioinformatics (INB), Population Genomics Node. [4] Institució Catalana de Recerca i Estudis Avançats (ICREA).

Epistasis in its broadest sense could be defined as the dependence of the outcome of a mutation on its genetic background. Mutations in functionally related genes could be selected jointly if they jointly affect the total fitness of individuals carrying them. Whenever this happens, the signature that selection may leave in patterns of genome variability might differ dramatically from the signature of selection when it acted upon a single gene. An excellent case-study is provided by ancestral and current human populations, which experienced significant changes in their selective pressures after leaving Africa to colonize the whole Planet. In this study, we studied derived human populations looking for evidence of recent positive selection acting upon pairs of protein-coding genes that are known interact within pathways (“interacting genes”). Evidence of recent positive selection events was investigated using Tajima’s D and other frequency-spectrum statistics. We found an excess of interacting pairs of genes with evidence of recent positive selection in derived populations that are not expected by chance when compared with the rest of the genome. Furthermore we found an excess of interacting pairs of genes with evidence of recent positive selection in derived populations when compared with African populations. These observations cannot be accounted for neither by (1) the demographic history of populations nor by (2) an excess of outlier genes in derived relative to African populations. Further analyses showed that the pairs of outlier interacting genes tend to be more central in the networks than the rest of outliers genes, and that this trend is stronger in derived than in African populations. Taken together, our results suggest that changes in selective pressures in human derived populations, instead of affecting only individual genes, promoted the simultaneous action of selection upon groups of interacting genes in a much greater degree than believed so far

EBV strain variation in different lymphoblastoid cell lines derived from 1000 Genomes Project individuals.

Gabriel Santpere Baró[1], Arcadi Navarro i Cuartiellas[1], Fleur Darré[1], and Mar Albà[2]

1 IBE (UPF/CSIC) Barcelona, Spain. 2 Research Unit on Biomedical Research (GRIB) ICREA-Municipal Institute on Medical Research (IMIM), Barcelona, Spain.

Multiple Sclerosis (MS) is a chronic neurological disease that causes a variety of symptoms and has a great impact in many first world countries. Medical, economical and social consequences of MS are very significant. Infection by Epstein-Barr virus (EBV) is a robust risk factor for MS, so there is a pressing need to clarify the contribution of genetic variants from different EBV strains. In addition, the relationship between MS, EBV and its human host presents interesting evolutionary aspects, which have been unexplored so far. Establishing phylogenetic relationships among strains from different human populations might enlighten the current patterns of distribution of EBV and EBV-related diseases and help understanding co-evolution between the virus and its host.

Most people in the world (~90%) are infected by EBV, which establishes permanently in B-cells. Here, we aimed to reconstruct EBV sequences extracted from B-cell-lines belonging to 743 different individuals which have been fully sequenced within the 1000 Genomes Project. We know for sure that at least one particular EBV strain (B95-8) must be present, since B-cells of subjects were transformed using this strain to obtain an steady source of DNA. However, B95-8 possesses an specific deletion in its genome which may allow us to distinguish between natural and artificial EBV strains.

We have developed tools for viral DNA variant calling using EBV full-genome sequences obtained by means of next-generation sequencing data. We have determined the presence of natural EBV only in 15 individuals (notably, 12 of them were Africans) and obtained a set of variants which are specific of these natural strains. We have observed very little genetic variation among individual B95-8 strains.

Why do we have bad multiple alignments: transcript selection as a major source of errors.

José Luis Villanueva-Cañas[1] and M Mar Albà[1,2]

1 Evolutionary Genomics Group, Fundació Institut Municipal d'Investigació Mèdica (FIMIM)-Universitat Pompeu Fabra (UPF), Barcelona, Spain. 2 Catalan Institution for Research and Advanced Studies, Barcelona, Spain.

The number of transcripts for each gene stored in the databases is growing continuously. These different transcripts are mainly the result of alternative splicing and multiple transcription start sites and they often result in different protein isoforms. In large-scale evolutionary analysis we often wish to automatize processes such as the alignment of coding sequences of orthologous genes or gene families for estimating evolutionary rates. Constructing all possible alignments using the different protein isoform combinations is often not feasible and greatly complicates any subsequent analysis. Therefore, first we need to choose one protein isoform to use for each gene. Which criteria should we use? By default, ENSEMBL, one of the most widely used genome databases, uses the transcript with the longest coding sequence (CDS) from each species to make multiple alignments (Vilella, A.J. et al., 2009). However, this may result in the alignment of sequences with very different length and including non-homologous regions (Toll-Riera et al., 2011). For this reason we have developed an alternative method called Protein ALignment Optimizer (PALO), based on the selection of the combination with the minimum CDS length difference.

In order to evaluate the performance of the different methods we gathered three orthologous genes sets comprising orthologous genes from 4 different species: mammalian (13.153 genes), vertebrata (5.551 genes) and eumetazoa (1.612 genes). For most genes there were several possible protein isoform combinations. We obtained alignments for all possible isoform combinations and identified the one with the highest percentage identity (BEST) for each gene in the dataset. We then quantified how often the combination containing the longest sequences (LONG), the combination selected by PALO and a randonmly picked combination (RAND), resulted in an alignment with the same scores as BEST. For this analysis we used python scripts and the multiple alignment program MAFFT (Katoh, K. et al., 2002). In the about 30% of cases in which LONG and PALO chose different transcript combinations, LONG selected the BEST combination in 16-19% of cases, PALO in 60-71% of cases and RAND 16-21% of cases.

Therefore the heuristics employed by PALO resulted in alignments with higher scores than when using the LONG approach. In addition, the alignment length was more similar to the BEST alignment. We also found that the use of the PALO method resulted in lower branch-specific non- synonymous to synonymous (dN/dS) substitution rate values, decreasing the number of false positives in positive selection detection (Jordan, G. et al, 2011).

Jordan, G., Goldman, N. 2011. The effects of alignment error and alignment filtering on the sitewise detection of positive selection. Mol. Biol. Evol. 2011 : msr272v1-msr272.

Katoh, K., Misawa, K., Kuma, K., Miyata, T. 2002. MAFFT: a novel method for rapid multiple sequence alignment based on fast Fourier transform. Nucleic Acids Res. 30: 3059-3066.

Toll-Riera, M., Laurie, S., Albà, M.M. 2011. Lineage-specific variation in intensity of natural selection in mammals. Molecular Biology and Evolution 28: 383-398.

Vilella, A.J., Severin, J., Ureta-Vidal, A., Heng, L., Durbin, R., and Birney, E. 2009. EnsemblCompara GeneTrees: Complete, duplication-aware phylogenetic trees in vertebrates. Genome research 19: 327-35.

INB Extended Web Services Registry

José M. Fernández[1,2], Jose M. Rodríguez[1,2], and Christian Blaschke[1,2]

[1]Structural and Computational Biology Programme, Spanish National Cancer Research Centre (CNIO)
[2]Spanish National Bioinformatics Institute (INB)

In bioinformatics, and in general life sciences, there are plenty of web services built following the different web service paradigms and bioinformatic web service variants: REST, SOAP, BioMOBY[1], DAS[2], BioMart[3], etc... Although there are already web service repositories like BioMOBY and web service catalogues like BioCatalogue[4] or DASRegistry[5], we feel that many of the existing solutions are not extensible enough to be applied to new web service paradigms. Even worse, in some cases the source code of the registry is not available, so it cannot be reused, improved or extended.

We present the design and implementation of the new INB web service repository that will allow us to document and manage the web services developed by different groups in the INB being open and extensible to the entire range of web service paradigms and variants used in bioinformatics. Due to the large web service portfolio the INB possesses, one of the main concerns when designing the INB extended web service registry was that web services from any paradigm should be describable at a certain level. As the new registry is strongly inspired by the BioMOBY repository, one requisite was that RDF triples should be a superset of the ones used in BioMOBY to make it backward compatible.

The INB Extended Web Services Registry has been built taking into account the learned lessons from BioMOBY web service repositories, which are documented using RDF triples serialized as RDF/XML. This documentation style allows the embedding of semantic annotations related to the described web services, so it can be enriched with new RDF triples using predicates and concepts from existing ontologies.

Furthermore, this extended repository uses a new set of predicates, which allow identifying and fetching log usage and web service health status in a standardized way.

[1] BioMoby Consortium et al. Interoperability with Moby 1.0--it's better than sharing your toothbrush! Brief Bioinform. 2008 May;9(3):220-31
[2] Dowell RD. et al. The distributed annotation system. BMC Bioinformatics. 2001;2:7
[3] Zhang J. et al. BioMart: a data federation framework for large collaborative projects. Database (Oxford). 2011 Sep 19;2011:bar038
[4] Bhagat J. et al. BioCatalogue: a universal catalogue of web services for the life sciences. Nucleic Acids Res. 2010 Jul;38(Web Server issue):W689-94
[5] Prlić A. et al. Integrating sequence and structural biology with DAS. BMC Bioinformatics. 2007 Sep 12;8:333

Development of unigene-derived SSR markers and a Data Base of PCR Primers associated to the agronomic traits in Wheat

Gaboun F[1], Udupa S[1,2], Ibriz M[3], and Solaymani A[3]

[1] National Institute of Agronomic Research (INRA Morocco), PO Box 415, Rabat, Morocco,
[2] ICARDA-INRA Cooperative Research Program, B.P. 6299, Rabat, Morocco.
[3] University Ibn Tofail, Faculty of Sciences, Kénitra, Morocco
CONTACT (1): F. Gaboun, Unit of Biotechnology, CRRA-Rabat; INRA Morocco.
E. mail: gabounf@gmail.com

Molecular markers are important biotechnological tools for the improvement of durum and bread wheat. These markers can be applied for genetic diversity analysis, tagging genes of agronomic traits such as biotic and abiotic stresses and quality traits. In this study, the objectives were to develop user-friendly database containing information on already published the tagged markers available for marker-assisted selection of wheat and to identify and develop additional genic microsatellite markers by screening Unigene database of wheat.

Information on the tagged molecular markers in wheat, their PCR primer sequences, PCR conditions and germplasm information were collected and user-friendly database was developed using MYSQL and Java languages. The data was stored in relational database management system.

For developing genic microsatellite, an automated process was developed by programming and using existing bioinformatic tools. This process was used for acquiring Unigene sequence data from GenBank, annotating sequences, identifying microsatellites and designing primers for those microsatellites. A total of 30553 microsatellite sequences were extracted from 40870 wheat unigene sequences downloaded from the GenBank. The identified microsatellite unigenes contained four different types of mono- (A, G, C, T), 12 di-, 59 tri-, 96 tetra-, 29 penta and 26 hexa-nucleotides repeats. The identified unigene-derived microsatellites mainly comprised of 1281 mono- (23%; after excluding poly A and the poly T), 1081 for di- (9,46%), 2961 tri-(53,32%), 171 tétra- (3,04%), 31 penta- (0,5%) and 27 hexa- (0,48%) nucleotides sequences. A total of 30553 Unigene-derived microsatellite markers were utilized for primer design for flanking regions of microsatellite motif and designed successfully 26,343 sequence-tagged microsatellite markers for wheat.

Key words: Molecular markers, durum and bread wheat, biotic and abiotic stress, quality primers, PCR conditions, candidate genes, database, unigenes, SSR, markers.

Towards a sustainable African Bioinformatics network for genomics in Africa

Ghazal Hassan[1, 2]

[1] *Pluridisciplinary Faculty of Nador, University Mohammed First, Nador, Morocco*
[2] *Laboratory of Genetics and Biotechnology, Faculty of Sciences, Oujda, Morocco*

Africa is a unique place to study gene-environment interactions and data on Africans are needed for better ancestral representation in genomic studies. However genomic and genetic research is still limited in Africa. To address Africa's health and science needs and limitations, the US National Institutes of Health and UK Wellcome Trust launched a Genomics Initiative termed Human Health and Heredity for Africa (H3Africa). H3Africa is intended to encourage a contemporary research approach by African investigators to the study of the genomic and environmental determinants of common diseases for improving the health of African populations. Specific goals of H3Africa include establishing collaborative networks of African researchers pursuing genomics-based, disease-oriented projects and creating or expanding infrastructure for genomics research. Within this H3 Africa Initiative, a central role is given to Bioinformatics Network. A Pan-African proposal for Bioinformatics networking to support African Health and Genomics has been prepared. This H3 Bioinformatics Network (H3ABioNet) is built around four activities that include training, data sharing, communication, and high-level analysis and development of tools customized for genomics research in Africa. Specific aims of H3ABioNet are setting up a core infrastructure, providing tools and bioinformatics solutions, promoting research partnerships, building local capacity, and improving data storage and management and data access. Other Important aspects of the H3ABioNet include long-term sustainability, monitoring and evaluation, institutional and national commitment and outreach beyond the H3Africa Consortium. H3ABioNet is a nodes-based organization with a central node to coordinate and administer the network and distributed nodes. Based upon existing bioinformatics capacity, participating nodes were classified into full, associate and development nodes. Initial H3ABioNet network comprises 22 nodes (10 full, 7 associate and 5 development nodes) from 16 African countries. The Network and its full nodes should work toward helping associate nodes build capacity in bioinformatics and strive for full node status. The network has a set of milestones to achieve within the 5 year grant period. This H3 integrated effort will likely help to alter the present balance between Africa and the developed world in genomics and bioinformatics.

.

Analysis of minimal metabolic networks through whole-cell *in silico* modeling of prokaryotes

Xavier, Joana[1], Rocha, Isabel[1]

[1] IBB-Institute for Biotechnology and Bioengineering / Centre of Biological Engineering, University of Minho, Campus de Gualtar, 4710-057 Braga, Portugal

Systems Biology has been gaining attention in the last few years [1] due to the new datasets that are being generated fast, but also to the novel bioinformatics potentialities that are being developed to deal with these data. In this work we explore the definition of a minimal cell, in the context of recently established systems biology tools by relating the minimal cell concept to the concept of last universal common ancestor (LUCA).

We have used genome-scale metabolic modelling and Flux Balance Analysis (FBA) [2] to predict critical reactions - reactions that cannot be cut off the network, otherwise the biomass produced would be zero - of four different prokaryotic *in silico* models. We used *Escherichia coli* (biotechnologically interesting and model organism), *Thermotoga maritima* (accepted as one of the oldest lineages of bacteria living today [3]), *Methanosarcina barkeri* (one archaea) and *Buchnera aphidicola* (an endosymbiont with minimal genome).

By calculating critical reactions across different species we aimed at identifying minimal networks' components. By comparing the set of critical reactions among the organisms studied, the final set of conserved essential reactions, common to all five networks studied was composed of fourteen enzymes, most of them part of the Tryptophan, Tyrosine and Phenylalanine, Vitamins & Cofactor biosynthesis, and nucleotide pathways. From those 14 enzymes, Chorismate synthase is probably the most central enzyme in the results of this study as it is central for the Tryptophan, Tyrosine and Phenylalanine pathway. Phylogenetic analysis have shown before that all chorismate synthases known are of monophyletic origin (descend all from the same ancestral sequence) [4]. It should however be referred that *Mycoplasma genitalium*, the species used more to study minimal life does not have this enzyme.

Based on these results we concluded that current top-down approaches based on parasite's minimal genomes fail to predict minimal life's essential features desired for biotechnological applications and also in the identification of LUCA's metabolic features. We also propose that a bottom-up approach, based on previous top down approaches from more datasets than the genome, is the best in future work for modelling and, possibly, constructing minimal cells.

[1] Kitano, H. *Science* **295(5560)** (2002) p. 1662-1664.
[2] Varma, A. et al *Journal of Theoretical Biology* **165(4)** (1993) p. 477-502.
[3] Battistuzzi, F.U. et al *BMC Evolutionary Biology* **4** (2004)
[4] Macheroux P et al *Planta* **207(3)** (1999) p. 325-334

Bioinformatics projects supporting Enquiry-based learning in High Schools

Isabel Marques[1], Paulo Almeida[1] and José Pereira-Leal[1]

[1]*Instituto Gulbenkian de Ciência, Oeiras, Portugal*

Sciences have become increasingly unappealing to younger generations, as is well illustrated by the dwindling numbers of students choosing science degrees for their university education. One way to motivate high school students to the processes of science discovery is to involve them as active participants in research activities. Here, we describe a pilot project testing the feasibility of using bioinformatics as a tool for enquiry-based learning in high schools.

A set of exercises were designed to be used by final year high school students to complement their Biology curriculum. The exercises, which ran for several classes, were framed in the form of a research project where the students had to find the answers to a series of increasingly complex questions. These questions demanded research using standard bioinformatics resources such as browsing databases, finding genes, protein families, genetic variability, etc. An innovative aspect of this project was the joint development between scientists, teachers and students, all collaborating on different aspects of the development, implementation and testing.

The pilot project started in 2007 and was developed by the Bioinformatics and Computational Biology Unit of the IGC in close collaboration with students and teachers from two Portuguese high schools (Miguel Torga, in Queluz and Quinta do Marquês, in Oeiras). Content development and implementation involved a three-way dialogue with permanent updating in response to the feedback of teachers and students. The implementation included teacher training by bioinformatics specialists with the dual goal of training the teachers in novel methodologies (bioinformatics), but also to prepare them to become teacher trainers themselves ("train the trainers").

Evaluation of the project indicates that students rated very positively their experience in following this project, regardless of their own career aspirations. Students and teachers were motivated by being exposed to new technologies and cutting edge areas of scientific research. Teachers considered bioinformatics and the training they received very useful in transmitting biological concepts. Future avenues of this project include the expansion of the contents, as well as the range of ages they can target, and the expansion to other schools.

Ingebiol: A flexible web interface generator for command-line applications

Darío Guerrero-Fernández[1] and M. Gonzalo Claros[1,2]

[2]Departamento de Biología Molecular y Bioquímica, Facultad de Ciencias, Universidad de Málaga, Málaga
[1]Plataforma Andaluza de Bioinformática-Centro de Supercomputación y Bioinformática, Universidad de Málaga, Málaga

Application with more features does not have to be necessarily the most used one. It is a known fact that sometimes users prefer to use a less effective application if it is easier to deal with it. Moreover, they can reject to use an application if they find it hard to use. This is one of the most important problems of software in bioinformatics: there are a lot of useful and effective applications available as command line tools without an user interface that are infra-used due to natural rejection to command-line interfaces.

Nowadays web applications are increasing their popularity because they are easy to use, they must be compatible with any operating system, do not require users to install additional software in their computers, and provide a decentralized access to the application which helps to exploit resources with effectiveness. Furthermore, new techniques (like AJAX) have been used to improve web application's interfaces and mimic a desktop user experiences.

With Ingebiol we pursue two different goals:

- Provide users with easy to use web interfaces for legacy command line applications with an optional and transparent jobs submission procedure for supercomputing facilities. Moreover, users will find consistent interfaces for different applications. Job management is also a built-in characteristic, and gives the user the opportunity to browse, download, or delete already sent jobs.
- Give developers an easy to use and flexible framework to develop web interfaces for their existent or future developments. Developers do only need to focus on the algorithmic problem without regarding the user interface.

At the same time, Ingebiol provides a REST web service for each configured command without additional developer intervention, being this an important addition for free.

It is important to note that Ingebiol is not a programming library. Is best to think about it as a web application that generates web interfaces from a few configuration files stored in text files. This approach allows developers to setup new interfaces for command line tools by writing some simple configuration files and placing them in a fixed location.

Ingebiol has been developed in Ruby on Rails (RoR), making use of AJAX techniques to give users the best experience possible. Ruby on Rails is a powerful web development framework based on the Ruby language. It can be deployed in any UNIX like platform (UNIX, Linux and Mac OS X).

MAIS-TB: An integrated web tool for molecular epidemiology analysis

Patricia Soares[1], Carlos Penha-Gonçalves[1], Gabriela Gomes[1], José B. Pereira-Leal[1]

[1] *Instituto Gulbenkian de Ciência, Oeiras, Portugal*

TB (tuberculosis) is today the second highest cause of death from infectious diseases worldwide, despite improvements in diagnosis, treatment and living conditions [1] . The emergence of MDR (multi drug resistance) and XDR (extremely drug resistance) is threatening to make TB incurable. [2]

There is growing evidence that the genetic diversity of *Mycobacterium tuberculosis* may have important clinical consequences. In that sense combining genetic, clinical and demographic data will allow a comprehensive view of the epidemiology of bacterial pathogens and their evolution, helping to explain how virulence and other phenotypic traits evolve in bacterial species over time. [3, 4] Hence to understand TB, an integrative approach is needed.

Some databases on tuberculosis and/or infectious diseases are available, although none of them provide the complete clinical and demographic picture, not allowing the understanding of the molecular mechanisms leading from strain genotype to clinical phenotypes. [5] Therefore we created MAIS-TB (Molecular Analysis Information System - TB), an informatics system, which integrates molecular analysis of MTB isolates, with clinical and demographic information. This system provides a new tool to access and identify associations between TB strain types and clinical and epidemiological characteristics of the disease. [6]

This system was developed around clinical and demographic information on portuguese patients and the corresponding molecular data. The isolates were genotyped for each one of the standard methods: SNP, MIRU, RFLP and Spoligotype.

In order to facilitate it´s use this web tool is totally automated. It generates reports, restructures graphical presentations and creates dendrograms. Through the MAIS-TB interface it's possible to insert and download data, to browse the database or search specific parameters. Individual trees for each method is available and a supertree combining all of them is also present.

Its architecture make it simple to adapt to other diseases.

The data is stored in a MySQL database and the web interface is based on the Django web framework.

[1] Millet, J et al. *ScienceDirect* (2009)
[2] Hershberg, R. et al. *PLoS Biology* (2008)
[3] Coscolla M et al. *ScienceDirect* (2010)
[4] Thwaites, G. et al. *Journal of Clinical Microbiology* (2008)
[5] Abadia, E. et al. *ScienceDirect* (2010)
[6] Kong, Y. et al. *Journal of Clinical Microbiology* (2006)

jORCA – Jumping to the Cloud

J. Karlsson[1], Oscar Torreño[1], Oswaldo Trelles[1]

[1]*Department of Computer Architecture, Málaga University, Spain*

Large scale genomic projects are producing massive data sets as a result of advances in high throughput technology. Parallel and -in particular- cloud computing appears as the most promising approach to manage and analyze such massive data. This approach requires a great community effort to adapt existing software used in bioinformatics to be able to take advantage of cloud resources. An alternative -and perhaps easier- approach is to attempt to achieve speedups by using map/reduce techniques with legacy (unmodified) applications [1].

However, regardless of approach, it is necessary to also provide end users with consistent and user-friendly interfaces for these applications. jORCA [2] is a desktop client able to efficiently integrate different types of web-services and repositories by mapping metadata over a general definition to support service discovery and to achieve flexible inter-communication between tools.

This abstract reports the adaptation of jORCA to work with web-services deployed on cloud architectures. Providing a web-service interface for cloud resources is not difficult and is well supported on most cloud architectures but using these web-services becomes complex due to the size of data and cloud architecture cost models. Typically, all data transfer in and out from the cloud is charged. Therefore, data transfer should be minimized and data sets re-used, as much as possible. Even more, transferring huge data sets should include functionality for resuming transfers etc. This implies a previous data upload step to a dedicated cloud data storage before service execution. Cloud services can read data from the data storage by receiving data references.

In order to support this, we have extended the metadata set of our service repository with annotations that indicate that input parameters to the services should be references to the cloud data storage, instead of data values. jORCA now also supports uploading big data to the Azure Blob Storage. Uploads can be resumed if connection is lost, thereby supporting large data. It is also possible to send these references to services deployed on the Azure Cloud. These services support submitting a job, quering status and finally retrieving data references to the Azure Blob Storage. Work in progress include supporting services deployed on other cloud architectures, in particular on Amazon EC2 and IBM SmartCloud.

Acknowledgments

This work has been partially supported by grant AT2009-0025 "Acciones Integradas-MICINN" and FP7-VENUS-C projects. OT is funded by RIRAAF-ISCIII project. OTS and JK are members of the GN5-National Institute of Bioinformatics, Spain.

[1] Daniel Ramet et al.; "Mr-Cirrus: Implementación de Map-Reduce bajo MPI para la ejecución paralela de programas secuenciales", Proceedings of JPAR2011, 2011.

[2] Victoria Martín Requena et al.; "jORCA: easily integrating bioinformatics Web Services"; Bioinformatics, 2010, Vol 26(4). pp 553.

Spanish National Bioinformatics Institute (INB): A technological platform to support the OMIC sciences and personalized medicine in Spain.

Allan Orozco[1]

[1]Spanish National Bioinformatics Institute, Spanish National Cancer Center (CNIO) Address: C/. Melchor Fernandez Almagro nº 3, Madrid 28029-(Spain), email:aorozco@cnio.es.

The National Bioinformatics Institute (INB) [1] is a technological platform of the Carlos III Health Institute [2], an autonomous organism dedicated to biomedical research and to provide scientific technical services in Spain. The INB is integrated by ten nodes and a central node that includes strategic collaborations with the Barcelona Supercomputing Centre (BSC) [3] and National Center for Genomics Analysis (CNAG) [4]. Its main mission is to provide scientific and technical support to the Bioinformatics and Computing Biology developments in Spain and to project its activities internationally (for example, the INB is processing its participation as the Spanish node of the future European Bioinformatics Infrastructure: ELIXIR [5]).

The INB efforts have been centralized in the strengthening of a solid and functional platform that facilitates the application and provides support of different bioinformatics methods specialized in projects coming from the OMICS areas and molecular translational medicine. Also, it has a consistent education division available, that seeks to transfer knowledge by techniques, processes and bioinformatics technologies that were developed since its creation, which is also very useful for several experts from the areas of biology and health sciences.

In functional and reference terms, the INB has designed, developed and implemented an architecture based on the following fundamental layers: users, accesses (resources, programming and GUIT/Web/Tools) and computing resources (servers). This infrastructure allows to dispose a series of services in bioinformatics for the storage, access and exploration of massive production biological data of diverse genomic projects. Also, it has available a wide technological catalogue of Web Services (more than 700) that have been built through diverse protocols such as: BioMoby [6], DAS Services [7] and Soap [8]. Therefore, the INB has available a general technical frame that provides generic solutions to different bioinformatics problems and that at the same time is sensitive to being adapted and extended in a modular way to new problems coming from the data analysis in NGS. Additionally, the INB maintains a collection of bioinformatics systems that include, among others: aGEM [9], BABELOMICS [10], BLASTXP [11], CONTEXTS [12], DNAlive [13], MODEL [14], IntOGEN [15], GeneID [16], SNPator [17].

Additionally, The INB have a select set of software packages (tools), methods, algorithms, platforms, frameworks, web applications, desktop applications, web portal and others that cover areas such as genomics, like Genomic Analysis: GEM Library [18], SECISalm[19], SymCurv[20], Selenoprofiles[21], APPRIS[22], CExonic[23], Pupasuite[24], SPLASH[25], Dot Plot[26], Cogdon[27], SYSNPs[28], CNViewer[29], ACD[30], GiTools[31]; Gene Expresion Analysis: Flux Capacitor[32], Adun[33], Qnorm[34]; DataBase and Data Integration: Biocreative Metaserver[35], Biodata-SF[36], Moby-Miner[37], ByoDyn[38]; Sequence Analysis and Function Prediction: FunCUT[39], TreeDET[40], Phylemon[41], PitchEye[42], Modlink[43]; Protein Function and Structure Analysis: ProStar[44], Firestar[45], FlexServ[46]; Clinical and Biomedical Applications: BIANA[47], Pmut[48]; Ontologies: OwlViewer[49]; Visual Bioinformatics, aGEM[50], 3D Electron Microscopy BenchMark[51]; Evaluations and Assessment: GOPHER[52]. Also, a system for the archiving and direct execution of workflows in Taverna format (IWWE&M (INB Web Workflow Enactor & Manager) [53], and a system for the creation of web services under a defined ontology and supervised documentation jORCA (Integrated System for the Execution of Bioinformatics Services) [54] and MowServ2 [55] and GWASPI [56] as a user friendly desktop software for the management and analysis of GWAS data. On the other hand, most of the applications and resources are dedicated to support and collaborate in diverse genomic projects nationally and internationally. Therefore, the INB is participating in a series of scaled projects that go from small actions and collaborative tasks to the support of projects of greater magnitude. For example, we can mention some projects such as: IMID-Kit [57], Open PHACTS [58], AquaGenomics [59], ESP-SOL [60], Aphid Base [61], e-TOX [62] and OLEAGEN [63] and others[1]**.**

Finally, the INB has a wide experience that opens a gateway and covers different fronts related to data bases, tools, systems and technologies that are directed to the necessities of Biologists, Biomedical and Bioinformatics professionals that require methods, algorithms and bioinformatics technologies with the object to resolve complex tasks in the analysis of massive biological data.

References

[1] INAB [http://www.inab.org/]
[2] ISCIII [http://www.isciii.es]
[3] BSC [http://www.bsc.es]
[4] CNAQ [http://www.cnag.cat/]
[5] ELIXIR [http://www.elixir-europe.org/]
[6] BioMoby [http://www.biomoby.org/]
[7] Biodas [http://www.biodas.org/]
[8] SOAP [http://www.w3.org/TR/soap/]
[9] aGEM [http://agem.cnb.csic.es/]
[10] BABELOMICS [http://babelomics.bioinfo.cipf.es/]
[11] BLASTXP [http://cgl.imim.es:8080/blastxp/]
[12] CONTEXTS [http://contexts.bioinfo.cnio.es/]
[13] DNAlive [http://mmb.pcb.ub.es/DNAlive/]
[14] Model [http://www.mmb2.pcb.ub.es/]
[15] IntOGEN [http://www.intogen.org/home/]
[16] GeneID [http://genome.imim.es/geneid.html/]
[17] SNPator [http://www.snpator.com/]
[18] Visual Omics [http://bioinfoinb.cnb.csic.es/VisualOmics/]
[19] SECISaln [http://genome.crg.es/software/secisaln/]
[20] SymCurv [http://genome.crg.es/SymCurv/documentation.html/]
[21] Selenoprofiles [http://big.crg.cat/services/selenoprofiles/]
[22] APPRIS [http://appris.bioinfo.cnio.es/]
[23] CExonic [http://cexonic.bioinfo.cnio.es/]
[24] Pupasuite [http://pupasuite.bioinfo.cipf.es/]
[25] SPLASH [http://mmb.pcb.ub.es/SPLASH/]
[26] DotPLot [http://chirimoyo.ac.uma.es/dotplot/]
[27] Cogdon [http://chirimoyo.ac.uma.es/codon/]
[28] SYSNPs [http://www.sysnps.org/]
[29] CNViewer [http://code.google.com/p/cnviewer/]
[30] ACD [http://www.snpator.org/public/new_login/index.php?show=other_applications/]
[31] GITools [http://www.gitools.org/]
[32] FLUX [http://flux.sammeth.net/]
[33] Adun [http://lavandula.imim.es/adun-new/]
[34] Qnorm [https://chirimoyo.ac.uma.es/qnorm/]
[35] BCMS [http://bcms.bioinfo.cnio.es/]
[36] BioData-SF [http://chirimoyo.ac.uma.es/biodatasf/]
[37] Moby Miner [http://inb.bsc.es/applications/java/mobyminer/moby_miner.html]
[38] ByoDyn [http://cbbl.imim.es:8080/ByoDyn/]
[39] FunCUT [http://ubio.bioinfo.cnio.es/biotools/FunCUT/]
[40] TreeDet [http://treedetv2.bioinfo.cnio.es/treedet/index.html]
[41] Phylemon [http://phylemon.bioinfo.cipf.es/]
[42] PitchEye [http://inb.bsc.es/applications/java/pitcheye/PitchEye.jnlp/]
[43] ModLink [http://sbi.imim.es/modlink/]
[44] ProStar [http://mmb.pcb.ub.es/proStar/]
[45] FireStar [http://firedb.bioinfo.cnio.es/Php/FireStar.php/]
[46] FlexServ [http://mmb.pcb.ub.es/FlexServ/]
[47] BIANA [http://sbi.imim.es/web/BIANA.php/]
[48] PMut [http://mmb.pcb.ub.es/PMut/]
[49] OwlViewer [http://bioinfoinb.cnb.csic.es/VisualOmics/OwlViewer/index_OV.html/]
[50] aGEM [http://bioinfoinb.cnb.csic.es/VisualOmics/aGEM/home.html/]
[51] 3DEM Benchmark [http://i2pc.cnb.csic.es/3dembenchmark/]
[52] GOPHER [http://gopher.bioinfo.cnio.es/]
[53] IWWEM [http://ubio.bioinfo.cnio.es/biotools/IWWEM/workflowmanager.html/]
[54] jORCA [http://chirimoyo.ac.uma.es/jorca/]
[55] MOWServ2 [http://chirimoyo.ac.uma.es/MOWServ2/]
[56] GWASPI [http://www.gwaspi.org]
[57] IMID- Kit [http://imid-kit.bsc.es/index.php/]
[58] Open PHACTS [http://www.openphacts.org/]
[59] AquaGenomics [http://www.aquagenomics.es]
[60] ESP- SOL [https://chirimoyo.ac.uma.es/espsol/]
[61] Aphid Base
[http://www.aphidbase.com/aphidbase/community_links/international_aphid_genomic_consortium/]
[62] e-TOX [http://www.e-tox.net/index.html]
[63] OLEAGEN [https://chirimoyo.ac.uma.es/oleagen]

List of attendees

Family name	Name	E-mail	Affiliation
Abdelaziz	Soukri	a_soukri@hotmail.com	Faculty of science University Hassan II
Abia	David	dabia@cbm.uam.es	CBM
Agirre	Eneritz	eagirre@imim.es	Universitat Pompeu Fabra
Aguilar	Daniel	daguilar@imim.es	Structural Bioinformatics Lab. GRIB. Universitat Pompeu Fabra
Aibar	Sara	saibar@usal.es	Centro de Investigación del Cáncer (CiC-IMBCC, CSIC/USAL)
Alastruey-Izquierdo	Ana	Anaalastruey@isciii.es	Instituto de salud carlos iii
Alba	Mar	malba@imim.es	GRIB (UPF-IMIM)
Alloza Anguiano	Eva	alloza_anguiano@yahoo.es	CIPF
Almeida	Paulo	palmeida@igc.gulbenkian.pt	Instituto Gulbenkian de Ciência
Alonso López	Diego	diego.alonso@usal.es	Centro de Investigación del Cáncer
Aloy	Patrik	paloy@irbbarcelona.org	ICREA / IRB
Althammer	Sonja	sonja.althammer@gmail.com	Universitat Pompeu Fabra
Alves	Renato	rjalves@igc.gulbenkian.pt	Instituto Gulbenkian de Ciência
Alves	Rui	ralves@cmb.udl.es	University of Lleida
Andres	Eduardo	eandres@cnio.es	CNIO
Andrio	Pau	pau.andrio@bsc.es	BSC
Arcas Mantas	Aida	aarcas@imppc.org	IMPPC
Aviles	Fransces Xavier	FrancescXavier.Aviles@uab.es	IBB / UAB
Baeza	Carlos	carlos.baeza@uv.es	Universitat de València
Barbadilla	Antonio	antonio.barbadilla@uab.cat	IBB / UAB
Barrera Burgos	Víctor	vbarrera@imppc.org	Institute of Predictive and Personalized Medicine of Cancer
Barturen	Guillermo	bartg01@gmail.com	University of Granada
Baù	Davide	dbau@cipf.es	Centro de Investigación Príncipe Felipe
Bautista Moreno	Rocio	rociobm@uma.es	Plataforma Andaluza de Bioinformatica
Bellora	Nicolas	nicolas.bellora@upf.edu	Regulatory Genomics, GRIB
Bemzekri	Hicham	bhicham538@gmail.com	Andalusian Platform of Bioinformatics (PAB). Universidad de Málaga
Benaicha	Soumia	benaicha.soumia@yahoo.fr	Faculté des sciences Oujda
Blaschke	Christian	cblaschke@cnio.es	Instituto Nacional de Bioinformática
Bleda Latorre	Marta	mbleda@cipf.es	CIPF
Bonàs	Sílvia	silvia.bonas@bsc.es	Barcelona Supercomputing Center
Bonet	Jaume	jaume.bonet@gmail.com	Universitat Pompeu Fabra
Briansó	Ferran	ferran.brianso@vhir.org	UEB - VHIR
Brunak	Soren	brunak@cbs.dtu.dk	CBS
Cano Castillo	Miriam	miriam.cano@uma.es	Universidad de Málaga
Capella Gutierrez	Salvador	scapella@crg.es	CRG
Carazo García	José María	carazo@cnb.csic.es	National Center for Biotechnology - CNB
Carnero Montoro	Elena	elena.carnero@upf.edu	IBE- (UPF-CSIC)
Casadio	Jascha	jaschacasadio@gmail.com	Universitat Pompeu Fabra
Cases	Ildefonso	icases@imppc.org	IMPPC

Family name	Name	E-mail	Affiliation
Castellano Esteve	David	david.castellano85@gmail.com	UAB
Castelo	Robert	robert.castelo@upf.edu	Universitat Pompeu Fabra
Chang	Jla Ming	chang.jiaming@gmail.com	CRG
Claros	M. Gonzalo	claros@uma.es	Universidad de Málaga
Codo Tarraubella	Laia	laia.codo@bsc.es	BARCELONA SUPERCOMPUTING CENTER
Conchillo Solé	Oscar	txino@bioinf.uab.es	IBB-UAB
Conesa	Ana	aconesa@cipf.es	CIPF
Contreras-Moreira	Bruno	bcontreras@eead.csic.es	Estación Experimental de Aula Dei / CSIC
Cuesta de la Plaza	Isabel	isabel.cuesta@isciii.es	Centro Nacional de Microbiología, Instituto de Salud Carlos III
Curado	Joao	joao.curado@crg.eu	Centre for Genomic Regulation
Dall'Olio	Giovanni Marco	giovanni.dallolio@upf.edu	IBE, Institut de Biologia Evolutiva, CEXS-UPF (Barcelona, Spain)
Daura	Xavier	Xavier.Daura@uab.cat	IBB / UAB
de la Peña	Santiago	spena@biomol-informatics.com	Biomol-Informatics S.L.
De Las Rivas Sanz	Javier	jrivas@usal.es	Fundación de Investigación del Cáncer (FICUS)
de Pedro Puente	Xavier	xavier.depedro@vhir.org	UEB-VHIR
Dias Xavier	Daniela	danidiasxavier@gmail.com	Universidad Complutense de Madrid
Díaz Gimeno	Patricia	patrizdiaz@gmail.com	CIPF
Djebali	Sarah	sarah.djebali@crg.cat	Centre for Genomic Regulation (CRG)
Dohm	Juliane	dohm@molgen.mpg.de	CRG
Dopazo	Hernán	hdopazo@cipf.es	CIPF
Dufour Rausell	David	ddufour@cipf.es	CIPF
Duran	Miquel	miquel.duran@irbbarcelona.org	IRB Barcelona
Engelken	Johannes	johannes.engelken@upf.edu	IBE (UPF-CSIC)"
Erb	Ionas	ionas.erb@crg.es	CRG
Esnaola	Mikel	imolina@creal.cat	Fundación CREAL
Esteve	Anna	anna.esteve@uab.cat	CRAG-UAB
Eyras	Eduardo	eduardo.eyras@gmail.com	Universitat Pompeu Fabra
Ezkurdia	Iakes	iezkurdia@cnio.es	CNIO
Faria	Rui	rui.faria@upf.edu	BioEVO, UPF, PRBB
Fernandes	Francisco	fjdf@kdbio.inesc-id.pt	KDBIO (INESC-ID)
Fernández	Leyden	leyden.fernandez@bsc.es	BARCELONA SUPERCOMPUTING CENTER
Fernández González	José María	jmfernandez@cnio.es	Centro Nacional de Investigaciones Oncológicas (CNIO)
Ferrer-Costa	Carles	carles.ferrer@gendiag.com	GENDIAG
Flores	Oscar	oscar.flores@irbbarcelona.org	IRB Barcelona - BSC
Fontanillo	Celia	cfontanillo@usal.es	Centro de Investigación del Cáncer (CiC-IMBCC, CSIC/USAL)
Fornés	Oriol	ofornes@imim.es	Universitat Pompeu Fabra
Freitas	Ana Teresa	atf@inesc-id.pt	INESC-ID/IST
Frias Moya	Leonor	lfiras@pcb.ub.cat	CNAG
Gabaldón	Toni	tgabaldon@crg.es	CRG
Gaboun	Fatima	gabounf@gmail.com	INRA Maroc
García	Javi	javigx2@gmail.com	Universitat Pompeu Fabra
García García	Francisco	fgarcia@cipf.es	Bionformatics and Genomics Department. CIPF

Family name	Name	E-mail	Affiliation
Garcia Jimenez	Beatriz	beatrizg@inf.uc3m.es	Universidad Carlos III de Madrid
Garcia-Alonso	Luz	lgarcia@cipf.es	CIPF
Gel Moreno	Bernat	bgel@imppc.org	IMPPC
Gelpi	Josep Lluís	jlgelpi@gmail.com	BARCELONA SUPERCOMPUTING CENTER
Ghazal	Hassan	hassan.ghazal@fulbrightmail.org	UMP
Gomes Dos Santos	Helena	hgomes@cbm.uam.es	Unidad de Bioinformatica CBMSO
Gomez-Puertas	Paulino	pagomez@cbm.uam.es	CBMSO (CSIC-UAM)
Gonzalez	Santiago	santiago.gonzalez@bsc.es	BSC
González	Juan Ramon	jrgonzalez@creal.cat	Fundación CREAL
Gonzalez-Porta	Mar	mar@ebi.ac.uk	EMBL-EBI
Götz	Stefan	sgoetz@biobam.com	BioBam Bioinformatics
Graña Castro	Osvaldo	ograna@cnio.es	Spanish National Cancer Research Centre (CNIO)
Guerrero	Dario	dariogf@scbi.uma.es	SCBI
Guigó	Roderic	roderic.guigo@crg.cat	Center for Genomic Regulation (CRG)
Guillen Montalban	Yolanda	Yolanda.Guillen@uab.cat	UAB
Guney	Emre	emreguney@gmail.com	Universitat Pompeu Fabra
Gut	Ivo	igut@pcb.ub.es	Centre Nacional d'Anàlisis Genòmica (CNAG)
Haenzelmann	Sonja	shanzelmann@imim.es	IMIM
Hartasanchez Frenk	Diego	diego.hartasanchez@upf.edu	IBE (UPF-CSIC)
Heath	Simon	scheath@pcb.ub.cat	CNAG
Higuera Cabañes	Clara	clarah@solea.quim.ucm.es	University Complutense Madrid
Himmelbauer	Heinz	heinz.himmelbauer@crg.es	Centre for Genomic Regulation
Hospital Gasch	Adam	adam@mmb.pcb.ub.es	INB-Instituto Nacional de Bioinformática
Huerta-Cepas	Jaime	jhuerta@crg.es	CRG
Jiménez	Natalia	natalia@cnb.csic.es	CNB-CSIC
Karathia	Hiren	hiren@cmb.udl.cat	University of Lleida
Karlsson	Johan	tjkarlsson@uma.es	University of Malaga
Kemena	Carsten	carsten.kemena@crg.es	Centre de Regulacio Genomica
Kostadinov	Ivaylo	kostadinov@icm.csic.es	Institut de Ciències del Mar, CSIC
Laurie	Steven	slaurie@imim.es	IMIM
Lehner	Ben	ben.lehner@crg.eu	ICREA / CRG
Lise Olivia	Andrieux	liseoliviax@gmail.com	Barcelona Supercomputing Center
Llabrés-Segura	Mercè	merce.llabres@uib.es	Univ. of the Balearic Islands
Lloréns Rico	Verónica	veronica.llorens@estudiante.uam.es	Universidad Autónoma de Madrid
Logares	Ramiro	ramiro.logares@icm.csic.es	ICM, CSIC
Lomas	Rodrigo	rlomas@cipf.es	CIPF
Macias	Jose-Ramon	jrmacias@cnb.csic.es	Spanish National Center for Biotechnology
Maietta	Paolo	pmaietta@cnio.es	CNIO
Marbà	Martina	freestym@gmail.com	CIPF
Marcet-Houben	Marina	mmarcet@crg.es	CRG
Marco	Santiago	smarco@pcb.ub.cat	CNAG
Marín	Manuel Alejandro	manuelalejandr.marin01@estudiant.upf.edu	Universitat Pompeu Fabra
Markaide	Pablo	ppablerass@hotmail.com	University of Basque Country

Family name	Name	E-mail	Affiliation
Marques	Isabel	imarques@igc.gulbenkian.pt	Instituto Gulbenkian de Ciência
Marques-Bonet	Tomas	tomas.marques@upf.edu	ICREA/IBE
Marti-Renom	Marc A.	mmarti@cipf.es	Centro de Investigación Príncipe Felipe
Martin-Garcia	Fernando	fmgarcia@cbm.uam.es	Centro de Biología Molecular "Severo Ochoa"
Martinez Lopez	Alfredo	amartinezl@uma.es	Universidad de Málaga
Martinez Marigorta	Urko	urko.martinez@upf.edu	IBE (UPF-CSIC)
Medina Castelló	Ignacio	imedina@cipf.es	CIPF
Mele Messeguer	Marta	marta.mele@upf.edu	IBE (UPF-CSIC)
Mercader	Josep Ma.	josep.mercader@bsc.es	BARCELONA SUPERCOMPUTING CENTER
Meziane	Iman	imanmeziane@yahoo.fr	Rabat Medical School
Miñarro	Antonio	aminarro@ub.edu	Dept. Statistics. University of Barcelona
Minguez	Pablo	pablo.minguez@embl.de	EMBL
Minoche	André M	andre.minoche@crg.eu	Center for Genomic Regulation (CRG)
Mir	Arnau	arnau.mir@uib.es	University of Balearic Islands
Mirny	Leonid	leonid@mit.edu	MIT
Montes	Iratxe	iratxe.montes@ehu.es	University of Basque Country
Moran	Federico	fmoran@bio.ucm.es	INB-UCM
Morreale	Antonio	amorreale@cbm.uam.es	Unidad de Bioinformática. Centro de Biologia Molecular Severo Ochoa.
Mosca	Roberto	roberto.mosca@irbbarcelona.org	IRB Barcelona
Moulay Hfid	Youssoufi	hafid.youssoufi@gmail.com	faculté des sciences oujda
Muñoz	Antonio	amunoz@uma.es	University of Malaga
Navarro	Arcadi	arcadi.navarro@upf.edu	ICREA / UPF-CSIC
Nogales-Cadenas	Rubén	ruben.nogales@cnb.csic.es	Centro Nacional de Biotecnología-CSIC
Novoa	Eva Maria	evahoop@yahoo.es	Institute for Research in Biomedicine
Oliva	Baldo	baldo.oliva@upf.edu	Universitat Pompeu Fabra
Oliveira	Ana Sofia	asfo@itqb.unl.pt	ITQB-UNL
Orellana	Laura	laura.orellana@irbbarcelona.org	Institute for Research in Biomedicine Barcelona
Orozco	Modesto	modesto@mmb.pcb.ub.es	IRB / UB
Pallara	Chiara	chiara.pallara@bsc.es	Barcelona Supercomputing Center
Panjkovich	Alejandro	alejandro.panjkovich@bioinf.uab.cat	Institute of Biotechnology and Biomedicine
Pantano	Lorena	lpantano@imppc.org	IMPPC
Paramonov	Ida	iparamonov@imppc.org	IMPPC
Pardo	Miguel Ángel	mpardo@cipf.es	CIPF
Pascual Montano	Alberto	pascual@cnb.csic.es	Centro Nacional de Biotecnología-CSIC
Pascual-García	Alberto	apascual@cbm.uam.es	Centro de Biología Molecular Severo Ochoa (CSIC-UAM)
Pegueroles Queralt	Cinta	cpegueroles@imim.es	FIMIM-UPF
Pereira	Rui	ruipereira@deb.uminho.pt	University of Minho - Centro de Engenharia Biológica
Pereira Leal	José	jleal@igc.gulbenkian.pt	Instituto Gulbenkian de Ciência
Pérez Cano	Laura	laura.perez@bsc.es	BSC-CNS
Petit Marty	Natalia Paola	natalia.petit@upf.edu	IBE (UPF-CSIC)
Planas-Iglesias	Joan	joan.planas@upf.edu	Structural Bioinformatics Lab. GRIB. Universitat Pompeu Fabra
Planes	Francisco J	fplanes@ceit.es	CEIT

Family name	Name	E-mail	Affiliation
Poglayen	Daniel	dpoglayen@imim.es	Universitat Pompeu Fabra
Porta Pardo	Eduard	eporta@imppc.org	IMPPC
Prieto	Carlos	carlos.prieto@unileon.es	Instituto de Biotecnología de León (INBIOTEC)
Pryszcz	Leszek	lpryszcz@crg.es	CRG
Puiggròs i Maldonado	Montserrat	puiggros.montserrat@bsc.es	BSC
Quintana	Adrian	aquintana@cnb.csic.es	CNB
Radó-Trilla	Núria	nrado@imim.es	GRIB (IMIM-UPF)
Ramet	Daniel	dramet@uma.es	Universidad de Málaga
Ràmia	Miquel	miquel.ramia@gmail.com	Universitat Autònoma de Barcelona
Repchevsky	Dmitry	dmitry.repchevski@bsc.es	BARCELONA SUPERCOMPUTING CENTER
Retamosa de Ágreda	Germán	german.retamosa@naudit.es	Naudit High Performance Computing and Networking S.L.
Reverter	Ferran	freverter@ub.edu	UB
Ribeca	Paolo	pribeca@pcb.ub.cat	CNAG
Riera Veiga	Sergio Arturo	sarv_ibz@hotmail.com	Mestral
Rodríguez	Jose Manuel	jmrodriguez@cnio.es	Spanish National Bioinformatics Institute (INB)
Rojas Mendoza	Ana María	arojas@imppc.org	IMPPC
Rosa Rosa	Juan Manuel	jm.rosa@sistemasgenomicos.com	Sistemas Genomicos SL
Roson Burgo	Beatriz	beatriz.roson@usal.es	Cancer Research Center (CiC-IBMCC, CSIC/USAL)
Rossello	Francesc	cesc.rossello@uib.es	UIB
Royo Garrido	Romina	romina.royo@bsc.es	BARCELONA SUPERCOMPUTING CENTER
Rubio	Angel	arubio@ceit.es	CEIT
Rubio Camarillo	Miriam	mrubioc@cnio.es	Spanish National Cancer Research Centre
Ruiz de Villa	M. Carmen	mruiz_de_villa@ub.edu	Universitat de Barcelona
Sanchez	Alex	aanchez@ub.edu	Universitat de Barcelona
Santoyo López	Javier	jsantoyo@bioinfomgp.org	Medical Genome Project
Santpere	Gabriel	gabrielsantperebaro@gmail.com	IBE (UPF-CSIC)
Sanz	Ferran	fsanz@imim.es	IMIM
Sebastian Yague	Alvaro	bioquimicas@yahoo.es	EEAD - CSIC
Serra	François	fserra@cipf.es	Centro Investigación Principe Felipe
Serrano	Luis	luis.serrano@crg.eu	CRG
Soares	Cláudio	claudio@itqb.unl.pt	ITQB / UPF
Soares	Patricia	psoares@igc.gulbenkian.pt	Instituto Gulbenkian de Ciência
Souilmi	Yassine	yassinesouilmi@gmai.com	FSR
Tabas Madrid	Daniel	dtabas@cnb.csic.es	Centro Nacional de Biotecnología - CSIC
Taly	Jean-Francois	jean-francois.taly@crg.eu	CRG
Tamames	Javier	jtamames@cnb.csic.es	CNB-CSIC
Tarazona	Sonia	starazona@cipf.es	CIPF
Torreño Tirado	Oscar	oscart@uma.es	Universidad de Málaga
Torrents	David	david.torrents@bsc.es	ICREA
Triviño Pardo	Juan Carlos	jc.trivino@sistemasgenomicos.com	Sistemas Genomicos
Trussart	Marie Jeanne	marie.trussart@crg.eu	CRG
Tur	Inma	inma.tur@upf.edu	Universitat Pompeu Fabra
Usié Chimenos	Anabel	ausie@diei.udl.cat	Department "Ciències Mèdiques Bàsiques" of University of Lleida (UdL)

Family name	Name	E-mail	Affiliation
Valencia	Alfonso	valencia@cnio.es	Centro Nacional de Investigaciones Oncológicas
Vavouri	Tanya	tvavouri@imppc.org	IMPPC
Vegas	Esteban	evegas@ub.edu	UB
Vilas Medina	Javier	javivm1@hotmail.com	Mestral
Villà-Freixa	Jordi	jordi.villa@upf.edu	Universitat Pompeu Fabra
Villanueva-Cañas	José Luis	jlvillanueva84@gmail.com	FIMIM-UPF
Xavier	Joana	joanarcxavier@gmail.com	University of Minho - Centro de Engenharia Biológica
Yruela	Inmaculada	i.yruela@csic.es	CSIC
Zhu	Ana	azhu@embl.de	EMBL
Zúñiga	Sheila	sheila.zuniga@sistemasgenomicos.com	Sistemas Genómicos S.L.

Our thanks to:

INB Spanish National Bioinformatics Institute

PORTUGUESE **BIOINFORMATICS**

Generalitat de Catalunya
Departament d'Economia
i Coneixement

Genoma España

biobam
BIOINFORMATICS SOLUTIONS

INBIOMEDvision

Organization support by:

ISCB Student COUNCIL
INTERNATIONAL SOCIETY FOR COMPUTATIONAL BIOLOGY

UNIVERSITAT POMPEU FABRA

cnag
centre nacional d'anàlisi genòmica
centro nacional de análisis genómico

prbb
Barcelona
Biomedical
Research
Park