ECCB 2014 Accepted Posters with Abstracts

J: Methods and technologies for computational biology

J01: Francesco Musacchia, Swaraj Basu, Marco Salvemini and Remo Sanges. Annocript: a flexible pipeline for transcriptome annotation that can also identify putative long non-coding RNAs

Abstract: Next generation sequencing technologies, specifically RNA-seq, promise to give a comprehensive portrait of the whole transcriptome in a given organism, even in the absence of a reference genome (Grabherr et al. 2011), thus optimized tools for the annotation are required. Furthermore, unambiguously identification of long non-coding RNAs (lncRNAs) is challenging since they are being discovered in almost all organisms (e.g. mammals (Cabili et al. 2011), fishes (Pauli et al. 2012), insects (Young et al. 2012), etc.). Current software to annotate transcriptomes, search for homologies with known proteins and domains, but they do not have the ability to identify putative lncRNAs (Conesa and Götz 2008; Schmid and Blaxter 2008, 8; Philipp et al. 2012; Koski et al. 2005). We developed Annocript: an automated pipeline to annotate large transcripome datasets and predict putative lncRNAs.

The pipeline is developed using PERL, R and MySQL and organized in modules. The first module generates a MySQL database to collect, map, organize and store information from several biological databases such as Uniprot (UniProt Consortium 2013), Conserved Domains Database (Marchler-Bauer et al. 2013), Enzyme Commission (Bairoch 2000), Pathways (Morgat et al. 2012) and GO terms (Ashburner et al. 2000). This task is performed only once. The second module uses BLAST+ (Camacho et al. 2009) to align query sequences against Uniprot proteins and CDD domains. The pipeline allows also to search for ribosomal and other short non-coding RNAs against a database of custom subset of the NCBI nucleotide and RFAM divisions (Burge et al. 2012). Finally, it executes Portrait (Arrial,Togawa and Brigido 2009), to calculate the non-coding potential, and Virtual Ribosome (Wernersson 2006), to extract the longest Open Reading Frame (ORF). The third module generates an user-friendly table containing the annotations: proteins, domains, enzyme, pathways, GO terms and longest ORFs. A final heuristic evaluates all the results to give a binary classification for the query transcript to be coding or lncRNA. A final module generates outputs in GFF3 format (for parsing and visualization) and HTML with statistics and charts. Fasta files are generated for coding, long non-coding transcripts and for the longest ORF translations.

Annocript is: flexible, because modules can be run independently and plugins can be added; fully customizable, in fact many programs parameters can be easily modified; optimized for fast computation, since it uses specific BLASTx parameters that speedup the search without any loss in sensitivity (Korf 2003) and our own parallelization of rpsBLAST; offline-software, because internet connection is needed only for generation of the initial annotation database; automatic, because it does not rely on any manual operation.

Annocript is distributed under the GNU General Public License and is freely available at https://github.com/frankMusacchia/Annocript.

J02: Jerome Mariette, Frederic Escudie, Philippe Bardou and Christophe Klopp. Jflow: A fully scalable Javascript workflow management system

Abstract: Building rich WEB environments aimed at helping scientists analyse their data is a common trend in bioinformatics. These applications are often specialized WEB portals or generic Workflow management systems (WMS). The first class provides multiple services and analysis tools in an integrated interface for a specific experiment or data type. Quite often these systems hide the processing steps in the back-office. The second class, for example Galaxy (Giardine et al., 2005), is mainly focused on workflow creation and provides a rather poor end-user interface but enable to combine tools and data sources as desired. We introduce jflow a fully scalable WMS that can easily be embedded in any WEB site, providing all WMS features and benefits to your project.

Jflow core is based upon the Makeflow (Albrecht et al., 2012) workflow engine and weaver (Bui, 2012), its Python API which is embedded within jflow. Using makeflow permits to run jflow under different batch systems such as Condor, SGE, Work Queue or a single multicore machine.

Adding a new component in the system requires to write a Python class inheriting from the Component class and to overwrite the process method wrapping the new tool. The class provides access to multiple concepts such as map/reduce or multimap in order to define the way the command line pattern should be applied to the input data. Multiple formats are available as input and output in order to force the workflow creator to combine components the right way. In the same way, writing a workflows consist of inheriting from the Workflow class and overwriting the process method. In this last method, the workflow is defined as the succession of components. Moreover, a property file is to be created to define the workflow parameters. It gathers all the information including the data type which will be checked from the provided interfaces. As an example, a date type will be displayed as a calendar in the graphical interface and the dd/mm/yyyy format compliance will be checked in command line mode.

All information about a workflows will be accessible from both jflow command line interface and its REST API. Thus, users can list available workflows and their states, run and monitor them. Accessing those functionalities from the command line interface can easily be done using the jflow command line. The same thing can also be done from a website integrating the jflow plug-in. To do so, jflow offers four modules to retrieve workflows information. As a jquery plugin, these modules are fully scalable and provide multiple methods and events to ease jflow integration. As example, the “click” event on a specific workflow triggers an event that can be listen to and used to build a workflow form wherever the web site designer wants it to be showed.

jflow http://bioinfo.genotoul.fr/jflow is a simple and efficient solution to embed workflow management systems features within any Web application.

J03: Jean Philippe Tamby, Rim Zaag, Jean-Paul Bouchet, Cecile Guichard, Philippe Grevet, Marie-Laure Martin-Magniette, Sébastien Aubourg and Véronique Brunaud. Evolution of the FLAGdb++ integrated environment for exploring plant genomes

Abstract: Comprehensive researches in genomics, post-genomics or systems biology involve the exploitation of large data sets. Therefore, databases provide key solutions to store and organize huge bulks of data, and ease the querying of information.

The FLAGdb++ information system, i.e. a core database coupled to a distributed user-friendly Java interface, was conceived to gather complete sequences of selected plant genomes and heterogeneous data coming from experimental and/or computer prediction approaches.

Using the genome as physical reference, any structural or functional annotation is merged and anchored to it independently. Hence, FLAGdb++ allows considering genes in large contexts like the chromosome, gene family, orthology group or co-expression cluster and functional network.

In addition to the four plant species (Arabidopsis thaliana, Oriza sativa, Populus trichocarpa and Vitis vinifera) already hosted by the database, two new genomes were made available in the latest release of FLAGdb++: Cucumis melo and Solanum lycopersicum.
The A.thaliana genome was updated to comply with the R10 annotation release of TAIR, and was re-annotated with original as well as expertised data that were previously obtained from collaborative works. New features are now easy to query and visualize using the interface. Particularly, 8,787 SNPs from the SolCAP project were mapped to the tomato genome, paving the way to genetic variation studies; 171,900 RNAseq contigs were integrated to improve gene discovery in the melon's genome. Moreover, it is now possible to display alternative gene models as they are predicted for 30% and 27% of the A.thaliana and C.melo loci respectively.
FLAGdb++ proposes a range of tools that allow to decipher functional connections by integrating various structural and functional features, and to facilitate translational tasks between plant species. For instance, specific interfaces have been developed to select a subfamily of transcription factor or repeat element, and also to filter Gene Ontology groups using their evidence code, mirroring the quality and origin of the classification. The orthology relationships tool shows the cross-links between the integrated genomes and can perform multiple alignments of protein sequences, a particularly powerful feature when inferring gene function and carrying comparative analyses.
FLAGdb++ is a powerful and complementary alternative to the web-based genome browsers for querying large data sets in genomics and post-genomics studies. It provides plant biologists access to a rich array of original genomic resources, for six complete genomes for now, hopefully for more in the future.

J04: Yuriy Vaskin, Francesco Venco and Heiko Muller. SMITH: managing NGS data and workflows in a sequencing facility

Abstract: As NGS-based methods become standard for studying bio-macromolecules, the sequencing facilities have to keep track of multiple requests and deliver high quality results shortly. The sequencing requests involve diverse characteristics of samples, projects, experiment types, indexes, data processing steps, etc. The laboratory information management systems (LIMS) are available to support the process of the NGS data delivery. SMITH (Sequencing Machine Information Tracking and Handling) is a LIMS for Illumina data with the integrated workflow subsystem.
There are high-quality tools for handling and processing NGS data in laboratories. A commercial tool BaseSpace (https://basespace.illumina.com/home/index) by provides a friendly user interface and extendable workflow subsystem. An open-source tool MISO [Davey et al. 2012] is a scalable system for metadata handling in an NGS facility. However, requirements of a particular sequencing laboratory are in constant change, so it is difficult to find a tool extendable enough to meet all of them, which motivates developing a new one.
SMITH consists of two parts: the metadata manager and the workflow subsystem. The metadata manager tracks the samples at each stage of sequencing, gives information about reagents in use, provides role-based access to the system and data search interface, helps to assemble flow cells, etc. The workflow subsystem performs automatic analysis of raw data (BCL files), alignment and post-processing, according to a type of an experiment. It produces FASTQ data, BAM files, etc. The workflow subsystem is extendable enough to embed new analysis protocols. These parts allow SMITH to cover the whole process from a sequencing request to routine-analysis-free data with minimum human interaction, which significantly reduce the delivery time and the error rate.
SMITH is a Java EE web-based application that runs on a Java EE server (like Glassfish) with MySQL at the backend. The workflow subsystem uses Galaxy[Goecks et al. 2010] and

UGENE Workflow Designer[Okonechnikov et al. 2012] for streamline analysis. Taking the metadata of a run, SMITH generates scripts that are run corresponding workflows and put data in folders according to the data structure of a sequencing facility.
Demo web site: http://cru.genomics.iit.it/smith
Source code: https://yuriy_vaskin@bitbucket.org/yuriy_vaskin/smith.git
License: MIT

J05: E'Krame Jacoby, François Le Fèvre, Ludovic Fleury, Sandrine Lalami, Audrey Lemaçon, David Vallenet, Guillaume Albini, Claudine Médigue and Claude Scarpelli. BIRDS, a rule based framework for automating generation and management of bioinformatics treatments

Abstract: Next Generation Sequencing technologies generate a high volume of data that needs to be automatically processed by amounts of combined bioinformatics tools. Existing systems dedicated to manage such workflows (e.g. Taverna, Galaxy) are mainly focused on design and execution but not necessarily on automation and parallelization. Here, we present a new Java API named BIRDS for Bioinformatics Rules Driven System. It is based on the JBoss rules engine Drools and is designed to automate the generation and management of bioinformatics treatments, applying business rules that are driven by resource availability.
The main advantage of BIRDS is its capacity to manage any source and storage of data called resources (flat files, databases…), since BIRDS does not require data formatting convention or metadata declaration. Thereby, the computational biologist only declares set of resources to be processed by workflows. Each workflow is made of several modules and is orchestrated by BIRDS's rules, which query resource availability, generate jobs and control their execution. BIRDS supports large-scale analysis through its multi-thread and cluster execution implementation. For each treatment, different job execution contexts could be defined from local to cluster parallelization using batch systems (e.g. LSF, SLURM). A dedicated database has been designed to store treatment configuration and job description which allows users to easily explore, monitor, check execution status and manually execute their jobs through a Web interface. The database also maintains an history of all treatments for traceability and keeps a list of consumed resources in order to avoid regeneration of already done treatments.
BIRDS is currently used in the SynBioWatch project, which develops an innovative NGS screening platform to automatically identify specific organisms in complex environmental samples. The platform is made of three main workflows: a pre-processing step which filters out sequence reads and removes all uninformative sequences; a comparison step where all reads are compared to a comprehensive sequence databank; an automated rule-based classifier to raise alerts for threat response teams. The end-user can visualize results via taxonomy and genome coverage reports. All workflows are made of several biocomputing modules that are managed directly by BIRDS, ensuring synchronization, robustness, control and traceability of job executions on HPC clusters.
BIRDS is much more than automation as it has been designed to be flexible and adaptable to any business constraint using the power of the Drools rule engine. It allows users to define business rules in a separate way and delegates all automation functionalities to BIRDS. These rules may act at several stages from command line building to pre and post job execution, including alerting and restarting jobs. BIRDS aims to be available in an open source mode (https://www.genoscope.cns.fr/projects/birds/).

J06: Oliver Horlacher, Frederic Nikitin, Davide Alocci, Julien Mariethoz, Markus Mueller and Frederique Lisacek. MzJava: an open source mass spectrometry library

Abstract: In order to provide bioinformatics support for new and experimental mass spectrometry (MS) experimental techniques, or to make better use of the information obtained from contemporary techniques, it is often necessary to write custom data analysis software. To make MS software development easier and faster we have developed MzJava, an open-source Java library that provides well engineered building blocks that are common to most MS data processing software.

MzJava was designed to be extensible, flexible and efficient so that it can be used to write research software and to analyze large data sets. MzJava provides algorithms and data structures for representing and processing mass spectra and their associated biological molecules, such as metabolites, glycans and peptides. The library contains algorithms to perform peak processing (e.g. centroiding, filtering, transforming), mass calculation, protein digestion, fragmentation of peptides and glycans and scoring of spectrum-spectrum and peptide-spectrum matches. For data import and export MzJava implements readers and writers for the commonly used data formats.

MzJava has been used to develop algorithms for positioning post translational modifications, clustering data independent MS/MS spectra[1], spectral library searches [2,3] and infrastructure code for EasyProt [4] a mass spectrometry software platform. MzJava is distributed under the AGPL v3.0 license and can be downloaded from http://mzjava.expasy.org.

[1] Pak, H. et al. (2013), Journal of the American Society for Mass Spectrometry, 24(12), 1862–71.

[2] Ahrne, E. et al. (2011). Proteomics, pages 1–11.

[3] Ahrne, E. et al. (2011). JPR, 10(7), 2913–21.

[3] Gluck, F. et al. (2013). Journal of proteomics, 79, 146–60

J07: Yannick de Oliveira, Olivier Sosnowski, Alain Charcosset and Johann Joets. BioMercator 4.0: A complete framework to integrate QTL, meta-QTL and genome annotation

Abstract: Compilation of genetic maps combined to QTL meta-analysis has proven to be a powerful approach contributing to the identification of candidate genes underlying quantitative traits. One of the most interesting properties of meta-QTL (or consensus QTL) is its confidence interval (IC) often shorter than IC of corresponding QTLs, decreasing the number of candidate gene to consider. As map compilation and QTL meta-analysis do not rely on genotyping raw data or trait measure, they can be easily achieved even if user holds maps from the literature or genetic databases.

BioMercator was the first software offering a complete set of algorithms and visualization tool covering all steps required to perform QTL meta-analysis. Despite several limitations the software is still widely used. We developed a new version proposing additional up to date methods and improving graphical representation of large datasets. As a major new functionality, user may now import sequence and genome annotations datasets within the software in order to display and mine functional annotation related to QTL and meta-QTL. BioMercator V4 is freely available from http://moulon.inra.fr/biomercator

J08: Yannick De Oliveira, Guy-Ross Assoumou-Ella, Johann Joets and Alain Charcosset. Thalia : A database dedicated to association genetics in plants

Abstract: Introduction: Diversity and association genetics studies lead to manipulate a large number of individual, lines, clones and/or populations. Moreover, emergence of high-throughput technologies for both genotyping and phenotyping generates a large amount of data. These need to be stored and managed in order to perform requests and organize datasets to conduct association genetics studies.

The Thalia database manages genetic resources, phenotyping and genotyping data, and also population structure information. Thalia enables data extraction in formats used by genetic association software.

Data Structure: The schema enables dynamic description of accession (an introduction within the collection) and seed lot types. Thus, users can describe accession types. Images can be linked to an accession.

This dynamic description is also available for markers. Thus, users can manages SNP, SSR, or other kind of marker. Some recent improvement ensure NGS data management. A (DNA) sample is characterized for a locus in a given experience (conducted by a person in an institute). One or more alleles are observed with given frequencies. Alleles are described following a referential used for the experiment. A correspondence option allows to bind heterogeneous data.

Phenotyping data are stored as expertised data relative to a seed lot observed in a given environment (average values or other statistical estimation). Raw data are stored in compressed files.

Classifications are expertised results concerning the assignation of a seed lot to classes of a population structure analysis [1]. All data in Thalia are managed in projects. A user can access to the data concerning projects in which he is involved. Some users can have an administrator status, which give them rights to insert data, and link data and users to projects.

Requesting and Analyzing Data: Each user can request and extract data concerning projects he is involved in. Genotyping and phenotyping data can be requested separately, but it is also possible to cross those data to extract information for association genetics studies or to interact with SniPlay [2] and GnPAsso databases. Classification results can also be requested.

A Google map viewer has been integrated to Thalia. For accessions associated with geographical coordinates, this viewer makes it possible to display classification, allele frequencies, trait or accession repartition on the map.

References

L. Camus-kulandaivelu, J-B Veyrieras, D. Madur, L. Combes, M. Fourmann, S. Barraud, P. Dubreuil, B. Gouesnard, D. Manicacci, A. Charcosset, Maize Adaptation to Temperate Climate: Relationship Between Population Structure and Polymorphism in the Dwarf8 Gene. Genetics 172, 2449-2463, 2006.

A. Dereeper, S. Nicolas, L. Le Cunff, R. Bacilieri, A. Doligez, J-P Peros, M Ruiz, P This, SNiPlay: a web-based tool for detection, management and analysis of SNPs. Application to grapevine diversity projects. BMC Bioinformatics 12:134, 2011.

J09: Jeongsu Oh, Chi-Hwan Choi, Soon Gyu Hong, Wan-Sup Cho and Kyung Mo Kim. CLUSTOM-CLOUD: In-Memory Data Grid-based software for clustering large scale 16S rRNA sequence data in the cloud environment

Abstract: Clustering 16S rRNA sequences generates operational taxonomic units (OTUs) to which individual sequences are assigned. Since the number of OTUs is the direct measure of microbial diversity, sequence clustering becomes the most important step when studying microbial ecology of a given environment. Even though many existing programs are available for clustering 16S sequences produced by next generation sequencing platforms, CLUSTOM was recently shown to be the most accurate. However, the original version of CLUSTOM takes a lot of time to cluster a large number of 16S sequences since its performance depends on the scale-up of CPU and memory resource of a single computing node. In order to overcome this limitation, we developed a new version named CLUSTOM-CLOUD that can be deployed in a cloud environment. By using the In-Memory Data Grid (IMDG) technique that has become popular recently in the industrial field of information technology but is not adopted yet in the field of bioinformatics, CLUSTOM-CLOUD was highly optimized to effectively use the resources of CPU and memory from distributed computing nodes and to support high availability. Therefore, CLUSTOM-CLOUD is an order of magnitude faster than the original version and enables to cluster over one million 16S sequences. Even though some of the distributed computing techniques such as Message Passing Interface (MPI) and MapReduce of the Hadoop ecosystem are available for processing large-sized DNA sequence data, the IMDG-based CLUSTOM-CLOUD has additional novel features in comparison to bioinformatics tools based on MPI or MapReduce: (I) rapid processing of large amount of data by allocating input data on IMDG that integrates random access memory of distributed computing nodes into a shared memory pool; (II) operating system or computing platform independent, supporting various computing environments such as personal computer, cluster and cloud environment like Amazon EC2; (III) provides highly scalable computing environment without any complicated installation or setting. These features distinguish CLUSTOM-CLOUD from existing cloud-based bioinformatics tools. CLUSTOM-CLOUD is written in JAVA and is freely available upon users’request.

J10: Katarina Truvé, Martin Norling and Erik Bongcam-Rudloff. SEQscoring: a web-based tool for interpretation and visualization of case control data from massive parallel sequencing (MPS) projects

Abstract: SEQscoring was designed to facilitate analysis and enable extraction of the most essential information from data produced in MPS projects. Its main goal is to help researchers locate the most likely causative mutations for a specific trait or disease. SEQscoring is supposed to be used after mapping and variant calling and accepts input data in some common formats including VCF. The emphasis is on data visualization, interpretation and ease of use. Cases and controls are compared in several ways and the output in form of graphs may be redirected to view in integration with data from the UCSC genome browser. SEQscoring also utilise the power of comparative genomics, by scoring all variants according to their degree of conservation. Variants in regions that are conserved between species are considered to be more likely to have a function and to play a role in phenotypic variation. The user can choose between a few different multiple alignments of species for scoring of conservation. Apart from visualization and comparisons of SNP and indel data, SEQscoring have some functionality to identify regions under purifying selection, and to locate larger structural variants like deletions and duplications. The intention is to help researchers extract a set of the most likely causative variants for further investigation. The SEQscoring tool is publicly accessible via the Web http://seqscoring.org/index.php. It has an intuitive interface and can easily be used by biologists, medical researchers, veterinarians as well as bioinformaticians. The version presented here is an enhanced upgrade with recent species assemblies, extended information content and improved functionality, since first published. (SEQscoring: a tool to facilitate the interpretation of data generated with next generation sequencing technologies EMBnet journal, 17.1 2011, pp. 38-45. DOI: http://dx.doi.org/10.14806/ej.17.1.211)

J11: Daniele Pierpaolo Colobraro and Paolo Romano. A new implementation of CABRI Web Services

Abstract: Information provided by Biological Resource Centers (BRCs) is of increasing interest for researchers.

It mainly consists in catalogues of well conserved and characterized micro-organisms that can be requested from BRCs.

IST Bioinformatics Web Services (IBWS) (1) have been deployed at the National Cancer Research Institute of Genoa, now IRCCS AOU San Martino IST.

They currently include three main groups, one of which make reference to CABRI, a web site for the integrated access to some European BRC catalogues (2,3).

In the context of the MIRRI project (4), three main improvements are being carried out:

i) extension of available information, by including some fundamental data that is available in some related databases,

ii) adoption of MCL (5) as a standard language for representing and exchanging catalogue information, and

iii) implementation of new WSs based on the REST standard.

Within IBWS, WSs have been developed both for each biological resources type included in CABRI and for all resources together.

CABRI WSs allow for the execution of a query either by name, by identifier or free text.

Two types of services were implemented:

i) searching for a specific feature (name and free text) and returning resource IDs, and

ii) searching for a resource ID and returning full records.

This approach will also be used for new WS.

We plan to extend WS to further information systems on microbial information which do not offer a programmatic interface.

We are currently considering the Global Catalogue of Microorganisms(6), Straininfo and some taxonomy related tools.

WS will be implemented by developing proper wrapper around the web interface of these systems because they do not offer programmatic access.

Access to WS will also be possible through new REST interface, that will complement the current SOAP interface.

The query and retrieval approaches will still be based on the idea that Web Services must at least reproduce the main features of standard web interfaces.

The output is also being revised. Beside the current simple ASCII result, an advanced format, based on the MCL language for representation and exchange of microbiological information, will be available. The new REST interface will be, of course, available as a standard web interface. A “human readable” version of the output will then be prepared too.

References

1. IBWS web site: http://bioinformatics.istge.it/ibws/
2. Romano et al. Applied Bioinformatics. 2005, 4(3):175-186. doi:10.2165/00822942-200594030-00002
3. CABRI web site: http://www.cabri.org/
4. MIRRI web site: http://www.mirri.org/
5. MCL project overview: http://www.straininfo.net/projects/mcl
6. GCM web site: http://gcm.wfcc.info/

J12: Rim Zaag, Jean-Philippe Tamby, Cécile Guichard, Zakia Tariq, Guillem Rigaill, Etienne Delannoy, Jean-Pierre Renou, Sébastien Aubourg, Marie-Laure Martin-Magniette and Véronique Brunaud. GEM2NET: From gene expression modeling to -omics network to discover Arabidopsis thaliana genes involved in stress response

Abstract: Although complete genome sequences for an increasing number of diverse organisms are available, the process of inferring function to genes is still at an early stage. Recent studies have estimated that 20% to 40% of the predicted genes have no assigned function for eukaryotic organisms whose genome is completely sequenced. At this point, there is a real need to overcome these limitations and improve our understanding of the functional relationships between genes and their products. Therefore, prediction tools and in silico analyses represent an inevitable alternative to step ahead.

Transcriptome data allow investigating the gene behaviors and co-expression studies have rapidly been considered as a way to identify sets of candidate gene modules. Generally co-expression is established by analyzing correlations between all the gene pairs from multiple microarray experiments collected from public repositories. Such approaches may suffer from both heterogeneity of data and the choice of the clustering method which is usually based on gene pairs.

To tackle these limitations, our analysis is founded on a large and homogeneous set of transcriptome data extracted from CATdb. This resource has the rare advantage to contain several thousands of microarray experiments obtained with the same technical protocol and analyzed by the same statistics pipeline. Furthermore, this resource is distinguished by about 5,000 genes that are not targeted by the commonly used ATH1 GeneChip® microarray of Affymetrix. We extracted 424 expression differences dealing with stress conditions and organized them into twenty stress categories representing various biotic and abiotic stresses. Concerning the co-expression analysis, for each stress type, it was performed with a model-based clustering method to propose a global point of view. Moreover, it offers a rigorous framework allowing the determination of the number of co-expressed gene clusters by a mathematical criterion. Genes were then assigned into the clusters according to their conditional probabilities. Through the twenty co-expression studies, more than 700 clusters of genes have been determined and the potential functions of these clusters are highlighted by several annotation analyses based on the Gene Ontology, sub-cellular localization, list of stress response genes, hormone genes, families of transcription factors.

By aiming at sharing the results with a broader community, all the analyses were stored and organized in a database to compose a dynamic Web resource, called GEM2NET, whose interface was developed using PHP and Javascript. For every type of stress, the clusters of co-expression are represented with original graphs that summarize the functional bias characterizing a group of interest at a glance. Moreover, integration of the Cytoscape Web module into the interface enables the visualization of external knowledge as protein-protein interactions or targets of transcription factors.

J13: Prachi Mehrotra, Vimlakany G Ami and Narayanaswamy Srinivasan. Classification of multi-domain proteins using CLAP server: case studies on proteins containing tyrosine phosphatases and SH3 domains

Abstract: The voids in the sequence and function space can be potentially filled using computational techniques that directly link the sequences to function. Our aim was to establish functional relationships between protein sequences through a robust classification scheme. Although several sequence comparison methods exist, many fundamental and computational limitations arise, especially in multi-domain proteins. The reason for the limitation is that the existing methods employ an alignment-based approach focusing on single domains in isolation. On the contrary, in the cellular environment, proteins are constantly regulated by the accessory and tethered domains, which contribute to the overall function of the multi-domain proteins. To this end, we have developed an alignment-free method to compute Local Matching Scores (LMS) using frequency distribution of fixed length patterns between two sequences.

Instead of globally aligning the sequences, LMS matches sequence patterns of length 5 amino acids and then scores them based on evolutionary substitution frequencies (BLOSUM62). Thus, it will accurately relate the sequences even if domain shuffling, duplication or circular permutation events have occurred. The normalized scores are hierarchically clustered using Wards' minimum variance method and the quality of clusters is assessed. As there are no gold standards for full-length protein sequence classification, we resorted to Gene Ontology and domain architecture based similarity measures to assess our LMS based Classification of Proteins (CLAP). CLAP is freely available as a web-server at http://nslab.mbu.iisc.ernet.in/clap/

Two test data-sets were chosen, one with tyrosine phosphatases and the other one with SH3 domain containing proteins from Pfam (sequence database version 27.0). The phosphatases family helped to ascertain whether our method performs equally well as the other alignment-based approaches, in the case of single domain proteins. SH3 proteins on the other side occur mostly as multi-domain in nature. Using CLAP, domain architecturally pure clusters with higher functional relevance were obtained. A sub–family level classification was obtained for the two families at particular distance cut-offs. We have also successfully shown that our method is ~ 7 times faster than alignment-based methods. Thus our method is instrumental in providing a biologically meaningful clustering of any given set of proteins, utilizing only the sequence information.

J14: Jason Williams. iPlant Collaborative: A Unified Cyberinfrastructure for the Life Sciences

Abstract: Research in biology increasingly depends on data-intensive methods and complex computational analyses that span a range of investigational domains - from genomics and transcriptomics, to phenotyping and ecology. The iPlant Collaborative (NSF #DBI-1265383) presents a comprehensive cyberinfrastructure (CI) that services a broad range of biological research questions by providing a unified platform for the storage, sharing, and analyses of large datasets – from genomes to image data, and beyond. Additionally, the variety of iPlant tools and access points cater to every level of user, from bench-biologists to bioinformaticians; removing many of the barriers researchers face in managing and sharing large data sets and conducting sophisticated computational analyses. iPlant tools are largely species agnostic and have grown to expand beyond the project’s initial focus of plant biology. iPlant CI is made of several interconnected platforms including: Discovery Environment iPlant's web-based platform for managing data, and running analyses is a scalable and extensible platform featuring hundreds of commonly used bioinformatics tools accessible through a simple sleek interface; Atmosphere iPlant cloud services for harnessing on-demand computational power and one-click access to a custom computing environment; Data Store iPlant's data storage system furnishes great flexibility and control over data; fast uploads/downloads, and terabytes of storage. iPlant Science APIs underlie much of iPlant CI, and can be accessed directly by developers. Computational resources include generous storage allocations as well as access to high-performance and cloud computing. iPlant platforms are extensible and customizable via application programming interfaces (APIs), RESTful services, and web-based systems for data access, tool integration, and analysis. Training and online learning materials make collaboration and people central to the cyberinfrastructure.

Funded by the U.S. National Science Foundation (#DBI-0735191), iPlant is driven by and freely available to the community. Our vision is focused on enabling researchers and educators to use cyberinfrastructure to solve problems that would otherwise remain insoluble. All users can register for a free iPlant account at www.iplantcollaborative.org

J16: David Brown, Rowan Hatherley and Özlem Tastan Bishop. HUMA: A web server and database for the analysis of genetic variations in humans

Abstract: HUMA (HUman Mutation Analysis) is a web server developed to allow researchers to find information linking genomic variation in humans to diseases. It also provides tools to model the 3D structure of a protein and to predict the structural effects of genetic variations.

The foundation of the HUMA web server is the database. It stores information about genes, transcripts, proteins, protein structures, diseases and variations in the proteins and transcripts.

It pulls this data from a myriad of data sources including the NCBI, PDB and UniProt. All information in the database is linked and any of these data categories can be used as a starting point when searching for information. For example, should a user query for information about a particular protein, the database will return all the information it has for that protein, as well as related structures, known variations in that protein, the transcript for the protein, the gene that encodes it, and any diseases that are known to be related to the protein. From there, a user can select one of the related variations. This will then filter the information for that particular variation, returning associated diseases if there are any as well as information from the other data categories.

The HUMA web server has been developed using the Django web framework. Data and functionality is exposed via a RESTful Web API. A modern, user-friendly, single-page web interface has been built on top of this API. As such, all the data and functionality that users can access via the interface can also be accessed programmatically via the web API. The web interface interacts with the API through Ajax calls, which negate the need to reload pages when sending requests to the server.

HUMA provides a number of tools that can be used for analysis. Firstly, the web interface uses GLMol for 3D visualization of protein structures. GLMol makes use of webGL to drastically improve the quality of 3D visualizations. Secondly, a homology modelling pipeline, which can model the 3D structure of a protein, has been developed and dubbed AutoModel. Lastly, HUMA provides a tool for predicting the structural stability of a protein. Used in conjunction with AutoModel, it will predict whether variations in a protein might cause instability. The functionality provided by these tools has also been made available via the web API. This will allow developers to include any and all functionality from the HUMA web server in their own scripts.

HUMA gathers data from a myriad of sources with the ultimate aim of providing researchers with everything they need in one place. HUMA provides an all-encompassing, logically designed web API that makes it easy for developers to include functionality from our server in their own scripts. We have built the HUMA interface on top of this API as proof of our commitment to maintaining it.

J17: Burçak Otlu, Sunduz Keles and Oznur Tastan. GLANET: Genomic Loci Annotation and Enrichment Tool

Abstract: High throughput sequencing technologies are routinely used for genotyping in genome-wide association studies (GWAS), profiling protein-DNA interactions, histone modifications DNA methylation, and for detecting copy number variations (CNV) across diverse phenotypes. These studies reveal a set of genomic regions of interest. Interpreting the relevance of these regions for the phenotype requires interrogating them within the context of diverse functional information. Currently there are several software tools that aim to annotate or perform enrichment analysis of a set of genomic loci. However, these tools are limited in terms of the type of loci and the enrichment analysis procedures which they can accommodate. We developed GLANET (Genomic Loci ANnotation and Enrichment Tool) for performing annotation and enrichment analysis of a given set of fixed or varying length loci. GLANET is an easy-to-run desktop application and utilizes predefined genomic coding regions (gene intron and exons, pathways) and noncoding regions including with a large collection of regulatory DNA element libraries from ENCODE. The readily available ENCODE libraries are obtained with next generation sequencing technologies (e.g., ChIP-seq, DNase-seq) exhibit systematic biases, which constrains regions of the genome that contribute to the libraries. To circumvent this shortcoming, we developed resampling-based enrichment test that accounts for such systematic biases. When we performed enrichment analysis on list of SNPs identified in Obsessive-Compulsive Disorder (OCD) GWAS, we have found novel pathways and transcription factors that are potentially linked to OCD.

J18: Dominik Lutter. A generalized additive model approach for high throughput screening approaches

Abstract: The availability of whole genome RNAi or shRNA libraries facilitated the use high screenings (HTS) for the study of gene function and the identification of new targets for drug development. Typically, these screens were performed in microtiter plates with each well, testing one RNAi on its genetic target, recorded in a specifically designed readout.

The most crucial step in the analysis if HTS data is the selection of hits, with a hit being a gene (or gene product) directly targeted by the corresponding RNAi. However, hit selection strongly depends on accurate data processing, which typically includes data normalization, correction of systematic effects (edge effects, plate-plate variances), noise reduction and quality checks. On top of this a number of different approaches for hit selection (Z-score, B-score, SSMD) have been developed. Generally, these steps of data processing form a string of individual, interdependent operations. Hence, the list of identified hits varies strongly on the selection of methods and the order of application.

Here we present a method based on generalized additive models. Based on the assumption that the data split up into different distributions, (hits and non-hits) our approach allows us to train a probabilistic classifier using plate and well specific intensity functions. The classifier can be trained utilizing whole screening data but higher power can be achieved by including control wells. We applied our approach on siRNA data, but the method is applicable to any kind of microplate performed HTS.

J19: Roland Barriot, Petra Langendijk-Genevaux, Yves Quentin and Gwennaele Fichant. Semi-automatic Validation of Genome-wide Reassembled Systems by Gene Prioritization through Weighted Data Fusion

Abstract: ATP-Binding Cassette (ABC) systems constitute a major super-family of systems present in all kingdoms of life, composed, in prokaryotic genomes, of two to five partners. We previously developed an annotation and reconstruction pipeline and maintain an online database ABCdb. The pace of new complete genomes releases requires tools to assist experts in the validation of the reassembled systems.

Here, we present a method for quality assessment of the reconstruction inspired by gene prioritization through genomic data fusion. The main innovations are (i) the weighing of the data sources used for prioritization and (ii) a genome wide approach for assemblies validation. The data sources used in this study and not included in the reconstruction pipeline are: expression profiles (GEO), annotations (Gene Ontology), interactions (STRING) and phylogenetic profiles.

Each data source leads to a gene pairwise similarity matrix used in the process of gene prioritization: candidate genes are scored according to their similarity to training genes. Applied to ABC systems, each reconstructed system constitutes a training set and all ABC genes (training genes included) are used as candidates. For each candidate a score is computed as the average similarity to the training genes. If the gene considered is part of the training genes then its similarity to itself is omitted when computing its score.

Each data source leads to a list of scored genes. To merge them, we perform a linear discriminant analysis (LDA) where each data source corresponds to a dimension and the similarity scores of a gene correspond to its coordinates in this multidimensional space. Through LDA, we obtain the weighted combination of dimensions that best separates the training genes from the others. This provides both the contribution of each data source, and

the fusion of the scores by applying the transformation. If the system was well assembled, the top genes are expected to be the training genes. Otherwise, the result might point to errors in the reconstruction.

Once all the systems of the studied genome have been evaluated, results are summarized as a directed graph in which nodes correspond to ABC systems and edges are added between systems when a gene from a system is also better ranked in another system. Three cases can occur. First, a node is isolated meaning that the corresponding system was well reconstructed. Second, there is an edge from a node to another (A -> B). In this case, we can consider that B is well reconstructed and its genes cannot be involved in A. As a result, A is also correctly assembled. Third, there can be cycles (A -> B -> C -> A or A -> B -> A) that reveal more complex situations that require expert manual intervention.

Applied on the 43 expertized ABC systems of at least two genes from Escherichia coli K12 MG1655, our strategy gives 18 isolated nodes, 19 nodes without cycles, and 6 nodes with cycles that correspond to ABC systems of the same subfamily.

J20: Sebastian Seitz and Tatyana Goldberg. Sorting the nuclear proteome using machine learning

Abstract: Subnuclear structures are associated with various nuclear processes. Hence proteins localized in these substructures are of great interest in understanding the interior mechanisms in the nuclei. Despite work on high-throughput methods, experimentally annotated proteins continue to be a minority. Even though predictions of subcellular localizations reach high levels of accuracy and coverage, the distinct substructures inside the nucleus are mostly not covered in these methods.

LocNuclei tackles this problem by using a profile based string kernel with support vector machines. In a two-step-approach it first distinguishes between sub- and supranuclear proteins. In a second step subnuclear localization in 14 distinct substructures, e.g. Nuclear Envelope or Nucleolus, is predicted. Using this approach high performance values were achieved, up to a mean AUC of 0.80 for subnuclear localization.

Depending on sequence-data for novel proteins only, LocNuclei creates a convenient method to assess intranuclear localization of newly discovered proteins and its results may be applied to other problems in biology.

J22: Aurélien Naldi, Pedro T. Monteiro, Denis Thieffry and Claudine Chaouiya. GINsim: a software tool for the modelling and analysis of logical regulatory networks

Abstract: The definition and analysis of large models of regulatory networks is currently a challenge in Systems Biology. The logical formalism is well suited to model a variety of biological systems, in particular when quantitative data are scarce. In this context, a model is defined by a Logical Regulatory Graph (LRG) encompassing regulatory components (nodes) and their interactions (arcs), along with logical functions defining, for each component, the target discrete levels of activity depending on the levels of the regulators. Model behaviour is usually represented in terms of a State Transition Graph (STG), where nodes denote logical states and arcs denote transitions.

GINsim is a software dedicated to the definition, simulation and analysis of logical regulatory graphs. To cope with the exponential grow of STG with the number of model components, several computational strategies have been implemented in GINsim, which either avoid the explicit generation of the STG by deducing properties directly from the model, or which significantly reduce its size.

GINsim uses Multi-valued Decision Diagrams to represent (multi-level) logical updating rules, which enabled the development of efficient algorithms (avoiding STG generation) for

the identification of all the stable states of a model, as well as to highlight regulatory circuits responsible for specific dynamical properties. GINsim further provides a reduction method that preserves most dynamical properties of the model. It also includes an algorithm to compact STGs on the fly, compressing them into a hierarchical graph, where the nodes represent connected sets of states or components, each symbolically represented by a decision diagram. Finally, GINsim supports model export in various formats, in particular towards NuSMV, thereby enabling the use of model-checking tools to verify temporal and reachability properties.
The GINsim team is involved in a collaborative initiative aiming to articulate different logical modelling approaches by developing common standards and an open software platform (CoLoMoTo, http://colomoto.org). In this context, a package of SBML Level 3 for the representation of qualitative models of biological networks has already been defined (SBML qual, http://sbml.org/Documents/Specifications/SBML_Level_3/Packages/Qualitative_Models_(qual)). Furthermore, a Java library has been developed to foster interoperability between logical modeling tools (LogicalModel, https://github.com/colomoto/logicalmodel). SBML qual has been implemented in LogicalModel to ease exchange of qualitative models between GINsim and other tools.
GINsim is currently used by various groups to model networks involved in cell fate decisions. It is freely available to academic groups along with a collection of logical models (http://ginsim.org).

J23: Pedro L. Varela, Pedro T. Monteiro, Nuno Mendes, Adrien Fauré and Claudine Chaouiya. EpiLog, a computational tool for the logical modelling of epithelial pattern formation

Abstract: The complexity of regulatory networks controlling cellular processes calls for the development of specific computational approaches. This complexity increases even more when cell-cell communication is involved, as is often the case in developmental processes such as epithelial pattern formation. Combining logical modelling, as implemented in the software GINsim (http://ginsim.org) and Cellular Automata (CA), we present EpiLog, a Java software for the qualitative simulation of epithelial patterning (http://ginsim.org/epilog).
In EpiLog, an epithelium is defined as a 2D grid of hexagonal cells. Each cell contains a Logical Regulatory Graph (LRG) that defines the molecular components, their interactions and behaviours. The latter are expressed in term of logical rules specifying the discrete level (of expression or activity) of each component, depending on the levels of its regulators. This LRG is specified, e.g. using GINsim, saved in the SBML qual format and loaded into EpiLog to be allocated to cells.
A LRG encompasses proper components, subject to regulatory effects from other components, and input components accounting for external influences. These can represent environmental cues or signals from neighbouring cells. Environment inputs, defined upon the whole grid, have constant values for each cell, depending on its position. In contrast, input components integrating signals from neighbouring cells evolve depending on an integration function. This is defined using a proper grammar based on the logical formalism, taking into account components secreted by neighbouring cells, numbers of these neighbouring cells and signal ranges.
A user-friendly interface supports the definition of these multicellular logical models, including initial conditions and colours assigned to the components for expression pattern visualisation. When the simulation is launched, successive states of the whole grid are displayed, combining the colours of the components (selected for visualisation) in each cell. The simulation ends whenever a stable pattern is reached, or when a maximum number of

steps has been completed. By default, all the components are updated synchronously at each simulation step. However, relying on prior biological knowledge, the user can define priorities, e.g. grouping faster components into a class first considered for update. Finally, beyond the study of the wild type case, EpiLog allows the definition of perturbations within user-defined regions of the epithelium.

In our poster, we introduce EpiLog and demonstrate its capabilities through a recently published model of the formation of dorsal appendages during Drosophila Melanogaster oogenesis.

Finally, we will discuss future extensions for updating strategies as well as current restrictions concerning cell proliferation, death and movement.

J24: Stephanie Le Gras, Serge Uge, Matthieu Jung, Ludovic Roy, Valerie Cognat, Frederic Plewniak, Irwin Davidson and Julien Seiler. GalaxEast: an open and powerful Galaxy instance for integrative Omics data analysis

Abstract: The exponential growth of high-throughput Omics data has posed a great technical challenge to experimentalists who lack bioinformatics skills and computing power. Moreover, integrative analysis of data from various sources is needed to provide biological insights into biological systems. We present hereby GalaxEast, an open and powerful public web-based platform for integrative analysis of Omics data (http://www.galaxeast.fr/).

GalaxEast is based upon Galaxy, one of the most popular bioinformatics workflow management systems, which is becoming a standard for sharing bioinformatics tools worldwide. As a Galaxy platform, GalaxEast aims at providing a large range of bioinformatics tools for the analysis of various types of Omics data. It supports reproducible computational research by providing an environment for performing and recording bioinformatics analyses.

GalaxEast is open to the entire academic scientific community with a guaranteed availability. Unlike other public Galaxy platforms, GalaxEast is designed for intensive computing. It is based on a high performance computing cluster composed of 24 cluster nodes with 400 CPU cores and 600Gb of RAM in total. The platform runs with the resource manager Slurm and is supported by 2 web servers and 2 job handlers to provide intensive computation and to handle a high number of simultaneous connections. The platform grants each user 50GB of storage and more storage can be granted temporarily if needed. Moreover, the available storage will be extended soon.

As a platform for integrative analysis of Omics data, GalaxEast is not dedicated to be a specialized Galaxy platform but a versatile one. Indeed it allows complete analyses from different sources of data from various research projects. It provides access to tools and algorithms devoted to Next Generation Sequencing (NGS) of epigenome (MACS), genomic sequence (GATK, Samtools), transcriptome (Cufflinks, HTSeq, TopHat) and statistical analyses (DeepTools, S-mart, DESeq). In the future, GalaxEast aims at implementing tools for other research fields (proteomics and imaging) to go toward a complete integrative analysis platform. De-noising methods for the analysis of spectroscopic data (urQRd) have already been implemented.

GalaxEast implements up-to-date standard Galaxy tools, plus bioinformatics developments from local research institutes. They cover different fields of research such as NGS data analysis for ChIP-seq (data manipulation, data conversion tools), and a genomic database querying tool to request the GEO Profiles database. Local developments are packaged and released through the open GalaxEast Toolshed.

GalaxEast also grants access to workflows developed for Omics data analyses either by the GalaxEast team or by the scientific community. Available workflows are dedicated to motif

search, repetitive element analyses or ChIP-seq data analysis, from data quality control to de novo motif detection.

J25: Y-H. Taguchi. Heuristic principal component analysis based unsupervised feature extraction and its application to bioinformatics

Abstract: Feature extraction (FE) is a difficult task when the number of features is much larger than the number of samples, although that is a typical situation when biological (big) data is analyzed. This is especially true when FE is stable, independent of the samples considered (stable FE), and is often required. However, the stability of FE has not been considered seriously. In this poster, we demonstrate that principal component analysis (PCA) based unsupervised FE functions as stable FE. Three bioinformatics applications of PCA based unsupervised FE: 1. detection of aberrant DNA methylation associated with diseases [1-2], 2. biomarker identification using circulating microRNA [3-4] and 3. proteomic analysis of bacterial culturing processes [5], are discussed.

In the first application, we have treated two examples: identification of genes with aberrant promoter methylation commonly associated with three autoimmune diseases [1] and identification of genes with genotype specific aberrant DNA methylation associated with Esophageal squamous cell carcinoma[2]. For both applications, we have successfully identified genes with significant probabilities.

In the second application, we have also treated two examples: identification of blood miRNAs discriminating between three liver inflammatory diseases and health controls [3] and identification of blood miRNAs discriminating 14 diseases and healthy controls. For these two applications. We have successfully identified sets of limited number (about ten) of miRNAs discriminating samples by 0.8 to 0.9 accuracies.

In the third application, we have applied PCA based unsupervised FE to culturing processes of bacteria, S. pyogenes that often causes life-threading diseases. In this application, our method successfully identified critical proteins in culturing processes of bacteria,

In conclusion, PCA based unsupervised FE is promising method which can be applied to wide range of bioinformatics applications.

References:

1) S. Ishida et al, (2014) Bioinformatic screening of autoimmune disease genes and protein structure prediction with FAMS for drug discovery, Protein Pept Lett. in press.(PMID: 23855671)

2) R. Kinoshita et al, (2014) Genes associated with genotype-specific DNA methylation in squamous cell carcinoma as candidate drug targets, BMC Syst Biol. 8(S1):S4.

3) Y-h. Taguchi and Y. Murakami, (2013) Principal Component Analysis Based Feature Extraction Approach to Identify Circulating microRNA Biomarkers, PLoS ONE, 8(6):e66714.

4) Y. Murakami et al, (2012) Comprehensive miRNA expression analysis in peripheral blood can diagnose liver disease PLoS ONE, 7(10):e48366.

5) YH Taguchi, Akira Okamoto (2012) Principal Component Analysis for Bacterial Proteomic Analysis, Pattern Recognition in Bioinformatics 2012, Lecture Notes in Computer Science, Vol. 7632, PP.141-152

J26: Xavier Prudent and Michael Hiller. Linking Phenotypes and Genomic regions: the Forward Genomics Approach

Abstract: Evolution has produced a great phenotype diversity among all species. Today we have many sequenced genomes but it mostly remains unclear which genomic regions are responsible for phenotype differences.

Previously, we developed a forward genomics method to match the divergence pattern of conserved genomic regions to the loss pattern of a given phenotype across a phylogeny, focusing on independent phenotype losses [1,2].

The original forward genomics method relied on finding a genomic region that is more diverged in all species where the phenotype is lost. While this simple approach works for some phenotypic differences [1], it leaves room for improvement.

Specifically, this simple approach only indirectly uses the fact that independent species have lost the same phenotype and it completely ignores that species evolve at different speeds, which affects the divergence of genomic regions.

Here, we present and compare new methods to detect genomic regions responsible for phenotype differences in a proper statistical framework.

Our new approaches take the evolutionary relatedness of the considered species (phylogeny) as well as their molecular evolutionary rates into account.

We use simulations that evolve entire genomes in silico to generate test sets under realistic parameters and present a performance comparison of these methods on these test sets.

[1] Hiller M et al. (2012): A “forward genomics” approach links genotype to phenotype using independent phenotypic losses among related species. Cell Reports, 2(4), 817-823

[2] Hiller M, Schaar BT, and Bejerano G (2012): Hundreds of conserved non-coding genomic regions are independently lost in mammals. Nucleic Acids Res, 40(22), 11463–11476

J27: Fotis Psomopoulos and Christos Ouzounis. Computation and visualization of ancestral pathway reconstruction and inference

Abstract: The emergence of metabolic pathways and their divergent evolution and diversification is the principal way by which organisms both obtain nutrients from the environment and discover new biosynthetic capabilities. The corresponding distribution of metabolic enzyme encoding genes across multiple genomes is surprisingly sporadic, following other, general patterns of genome conservation (PMID:15681613). The study of evolutionary relationships between metabolic pathways and their variants across genomes represents an emerging area of research that attempts to delineate metabolic capabilities encoded in the corresponding genome sequences, given the complexity of processes such as horizontal gene transfer (PMID:19266023). Moreover, the predicted role of a pathway can be evidenced by its evolutionary context in the form of phylogenetic profile signatures as well as the comparison to either functionally equivalent or evolutionarily divergent pathways.

The investigation of evolutionary histories including ancestral reconstruction has shifted from individual genes, to genomic structural segments, and more recently to biochemical functional modules. The inference of functional modules such as biochemical pathways is highly challenging as it relies not only on the detection of structural components of genomes but also of the corresponding enzymes and reactions, a much more error-prone process. Various approaches have been proposed previously, varying in detail, complexity, performance, degree of automation and usability. Based on alternative representations and adaptive thresholds (PMID:23341912), we have developed a robust and efficient algorithm that performs ancestral pathway inference across multiple genomes. These pathway ancestral states are computed on the basis of parsimonious criteria (PMID:12874054), phylogenetic profiling and algorithms that rely on these approaches (Psomopoulos et al. 2014a). To support the algorithmic development with interactive user feedback, we have also developed a visual representation of ancestral inference by means of a Cytoscape plugin (Psomopoulos et al. 2014b).

The implemented application exploits the inherent scalability of fuzzy phylogenetic profiles and is easily adapted for a high performance environment, such as the European Grid Infrastructure. We have obtained some preliminary results which have been validated against

established observations on specific sets of pathways/genomes.

Acknowledgements: We thank the European Grid Infrastructure and its supporting National Grid Initiatives for providing the technical support, computing and storage facilities. This work has been partly supported by the FP7 Collaborative Project Microme (grant agreement #222886-2), funded by the E.C. We thank Laura Dadurian for assistance.

References

Psomopoulos, F.E., van Helden, J., and Ouzounis, C.A. (2014a), submitted manuscript
Psomopoulos, F.E., Vitsios, D.M., and Ouzounis, C.A. (2014b), submitted manuscript

J28: Jörgen Brandt, Marc Bux and Ulf Leser. Cuneiform - Parallel Execution of NGS Workflows

Abstract: The storage and analysis of next-generation sequencing (NGS) data constitutes a major challenge for current computational infrastructures. Scientific workflow management systems have been proposed as a means to facilitate the design, execution, and iterative refinement of complex analysis processes as typically encountered in NGS data processing. In this application field, the sustained trend towards producing more data at lower costs implies the necessity for distributed workflow execution.

We present Cuneiform: A functional workflow language with a focus on distributed execution and interoperability. Tasks in this language can be defined in several programming languages to facilitate the reuse of existing software libraries and command-line tools. Furthermore, Cuneiform allows the user to exploit data- and task-parallelism. Like GATK's QScripts Cuneiform comes as a textual workflow language. In addition, Cuneiform combines the interoperability of scientific workflow systems like Galaxy or Taverna with the parallelism of large-scale data processing frameworks like Hadoop.

We illustrate the usefulness of Cuneiform as a workflow specification language by the example of a variant-calling workflow. We demonstrate how the workflow can be expressed in Cuneiform and how data-parallelism is exploited to obtain analysis results fast even from large data sets.

J29: Jasmin Straube, Alain Dominique Gorse, Bevan Emma Huang and Kim-Anh Lê Cao. A linear mixed model spline framework for analyzing time course ‘omics’ data

Abstract: Motivation: Recent advances in technology have enabled the quantification of thousands of molecules at all functional levels within an
organism. Moreover, it is now feasible to study not just a single time point, but a series of snapshots in order to reveal an individual’s
response to perturbation, developmental processes or the circadian cycle. However, ‘omics’ data derived from time-resolved experimental
designs are not only high-dimensional, but will also often have missing values and be noisy, all of which have adverse effects
upon analysis. Hence, revealing the underlying molecular response patterns requires reduction of sample dimension while addressing
subject-specific noise.

Results: We present a unified approach to analyse time-resolved ‘omics’ data which enables a user-friendly and convenient framework for data pre-processing and exploration. A data-driven iterative approach using linear mixed effect model splines and their derivatives was developed to model expression profiles while taking subject-specific variability into account. We show that modelling before clustering is beneficial to increase the biological relevance of findings. Additionally, we developed tests for differential expression based on the linear mixed effect model spline framework. Through simulation, we demonstrate the increased

sensitivity of our approach relative to common alternatives. Moreover we applied the methods to a real data set and obtained biological meaningful insights consistent with previous published studies.
Availability: The R package is available on the CRAN server. A webinterface is in progress.

J30: Svetlana Artemova, Mael Bosson, Jocelyn Gate, Sergei Grudinin, Leonard Jaillet, Marc Piuzzi, Petr Popov and Stephane Redon. SAMSON: Software for Adaptive Modeling and Simulation Of Nanosystems

Abstract: Modeling and simulation has become an essential tool for analyzing and designing complex molecular systems and numerous high-quality software tools have already been developed. Unlike most existing software, though, SAMSON tightly integrates modeling and simulation to aid in the analysis and the design of molecular systems. For instance, an interactive quantum chemistry module (ASED-MO level of theory) allows users to build and edit structures while interactively visualizing how the electronic density is updated. SAMSON also provides interactive flexing and twisting tools for large-scale flexible deformations of macro-molecular assemblies (e.g. proteins) with a few mouse clicks which, for example, may be used in docking and fitting tasks. Interactive virtual prototyping of hydrocarbon systems may also be used to edit and constrain graphene sheets, nanotubes, etc.
This integration is made possible via SAMSON’s data graph, which contains all information on models and simulators. Nanosystems are described through five types of models: structural models (geometry and topology), dynamical models (degrees of freedom), interaction models (potential energy, forces, electronic structure), visual models (graphical representations) and property models. SAMSON’s data graph relies on a signal-based system that may be used to develop adaptive algorithms. For instance, an adaptively restrained state updater may control, at each time step of a simulation, which degrees of freedom should be updated in a dynamical model [1]. In turn, an interaction model may request from a dynamical model the list of positions that have been updated since the last simulation step, to incrementally update the potential energy, forces and electronic structure [2, 3, 4].
SAMSON has an open architecture: the provided Software Development Kit allows developers to extend SAMSON’s functionality by developing SAMSON Elements, e.g. new interaction models, editors (e.g. procedural generators), apps (such as the DockTrina app for docking protein trimers [5]), wrappers or interfaces to existing software, connectors to web services, etc. SAMSON will be made available soon through SAMSON Connect, a website which makes it easy for developers and users to distribute and install SAMSON Elements(http://www.samson-connect.net). SAMSON will be free of charge for academia.

[1] S. Artemova and S. Redon, Physical Review Letters, 109:19, 2012
[2] M. Bosson et al, Journal of Computational Physics, 231:6, 2012
[3] M. Bosson et al, Journal of Computational Chemistry, 34:6, 2013
[4] R. Rossi et al, Bioinformatics, 23:13, 2007
[5] P. Popov et al, Proteins, 82:1, 2014.

J31: Nicolas Sapay, Ghita Rahal and Artem Khlebnikov. Multi-omics data integration platform in public private partnership

Abstract: Life sciences are currently undergoing a revolution. The improvement of information technologies and laboratory technologies triggers the generation of large amount of data from biological studies. Heterogeneity and complexity of the data collections prevent any transversal use and generates large computing costs to exploit them efficiently. Every community works with its own data scheme and computing models.
BIOASTER a novel French Technology Research Institute has built a partnership with the

CNRS National Computing Center (CC-IN2P3) to develop a computing model customized to the needs of the biosciences community. The partnership will benefit from the large experience acquired by CC-IN2P3 in the domain of High Energy Physics for Petabytes data experiments with grid-distributed infrastructures. CC-IN2P3 is a Tier-1 for LHC experiments at CERN.

BIOASTER and CC-IN2P3 will make use of the new generation cloud computing to address the specificities of the biological data creating dedicated core facility. This facility will provide high throughput data storage and powerful computing resources while customizing the needs to each category of usage. Cloud technology was chosen as it can provide the needed flexibility. Existing commercial clouds could be used, but they might not address the privacy and confidentiality required by biology data, and their cost might become quickly prohibitive.

In parallel, BIOASTER is building its scientific information system :

– to integrate the raw data between the BIOASTER’s core facilities and the partners ;
– to track the biological samples and the associated data in clinical and pre-clinical studies ;
– to share and visualize the data between partners.

A first brick of the information system will be the deployment of TranSMART on the cloud, as a tool to manage, share and explore clinical studies and the associated transcriptomics data. The next objective is to open this resource to other –omics disciplines, paving the way toward multiomics analyses.

J32: Jonathan Mercier, Alexandre Renaux, David Vallenet, Adrien Josso, François Lefèvre, E'Krame Jacoby Ayari, Guillaume Albini, Aurélie Genin-Lajus, Claude Scarpelli and Claudine Médigue. The MicroScope platform: from data integration to a rule-based system for massive and high-quality microbial genome annotation

Abstract: The emergence of the Next Generation Sequencing (NGS) generates an incredible amount of genomes, whereas curation efforts to annotate them tend to decrease despite some community initiatives. To ease this manual process, we develop the MicroScope platform: an integrated environment for the annotation and exploration of microbial genomes (https://www.genoscope.cns.fr/agc/microscope). It is made of three major components: (1) a management system to store and organize biological knowledge in relational databases, (2) a production system to organize and execute workflows, (3) a visualization system for expert analyses and data curation through a Web interface.

The Prokaryotic Genome DataBase (PkGDB) central model is enriched by the integration of numerous public databases collecting different types of biological entities. To support continuous data integration and reconciliation of these external resources, we designed the Galileo application (http://galileo.genoscope.cns.fr) based on a Model Driven Architecture. Its model manages the integration of several releases of the same biological resource and ensures unicity of biological objects from different resources with the use of internal business keys based on their key properties.

The MicroScope production system orchestrates about 25 workflows, which combine various bioinformatics software. To be able to manage billions of jobs in a HPC system we are migrating to a new API called BIRDS (BioInformatics Rules Driven System) and developed at the Genoscope. BIRDS is based on the Drools framework and provides a common environment for business rules and resource-driven workflows to automate bioinformatics treatments.

One important goal is to ease the human interpretation of genomic data in the light of predicted functions and biological processes (e.g. metabolic pathways). We are working on an explicit representation of the biological knowledge and on algorithmic tools designed to automate the biologist reasoning within the MicroScope platform. A first implementation of

such deductive reasoning has been implemented in the HERBS system through a collaborative project between INRIA and SIB institutes. A rule-based expert system, named Grools, is currently under development. This tool should improve predicted molecular functions, completeness and consistency of genome knowledge in the MicroScope platform.

J33: Óscar Torreño Tirado and Oswaldo Trelles. Easily registering bioinformatics services metadata

Abstract: Background: Web Services (WS) are the preferred way to provide bioinformatics services to the scientific community, facilitating remote and universal access. Unfortunately, due to the diversity and abundance of available services, end users were not able to discover and execute services due to being unaware of their existence. Consequently, the concept of WS repositories arose. These repositories store the services' meta-data (parameters, data types, documentation, etc.) in a centralized way. Currently there exist a number of these meta-data repositories in the bioinformatics field, with BioCatalogue (Bhagat, J. et al. 2010) being one of the most representative. However, it is still somewhat difficult to register new services in an uniform and easy way.

Methods: We have developed Flipper, a desktop application written in Java that is able to register services' meta-data, and package and deploy the executable of a new service.

Flipper uses MAPI (Ramirez, S. et al. 2011) as base software. This API has its own internal repository representation and can easily adapt the mapping of the diverse catalogs' information to the internal representation at any time, as well as define new mappings for new repositories.

Flipper manages the registration of:

1. Functional categories: for hierarchical organization of services.
2. Data types: to define the type of the service's parameters facilitating their interconnection.
3. Services: describing the service thereby facilitating its discovery and usage.
4. Namespaces: representing the domain to which the data is referring.

For deployment, Flipper provides automatically generated code templates that can be fine-tuned by the service provider.

Results

Flipper has been tested for functional categories, data types, registration of services and service deployment in the following repositories:

1. Spanish Institute of Bioinformatics (INB)
2. Advanced Clinico Genomics Trials on Cancer (ACGT)
3. Educational Software (EduSoft)

Conclusions: In this document we present a platform independent application that registers and deploys WS, facilitating their discovery, organization and execution. The deployment feature is important because it is usually not present in state-of-the-art software (i.e. BioMOBY Dashboard). The application works with multiple repositories, and can adapt to new ones without any change in the code. The reason is that we use MAPI as middleware, delegating it the required changes. These changes will be reflected not only in the registering application, but also in other software based on MAPI (i.e. jORCA (Martín-Requena, V. et al. 2010)).

Although the repositories unify the documentation of the services, it is still necessary to validate the provided information to ensure the quality of these repositories, an important requirement in large infrastructures such as Elixir. The presented application addresses this problem, helping service developers to complete all the documentation.

J34: Ryohei Suzuki, Daisuke Komura, Kazuki Yamamoto and Shumpei Ishikawa. MOLding: Gesture-based Interactive Molecular Dynamics for Protein Structure Manipulation

Abstract: Molecular Dynamics (MD) is a powerful tool for investigating complex biochemical reactions such as receptor-ligand bindings. MD simulates molecular behavior by solving equations of motion based on calculated potentials between atoms in the system, and provides various insights for understanding biochemical reactions. Interactive Molecular Dynamics (IMD) is a derivative technique of MD that permits manipulating molecules in molecular dynamics simulations by allowing user to apply external forces to atoms in a simulated system. While IMD is widely used in biochemical researches, present IMD implementations provide very limited means for specifying external forces for atoms, and it is still difficult to manipulate large structures of molecules (e.g. tertiary structures of proteins) in IMD systems.

We implemented a novel input method for IMD named “MOLding” that enables the users to use their hand gestures for manipulating molecular structures effectively, by extending the existing molecular visualization system, VMD (Humphrey et al., 1996).

If a user put his/her hands in front of a display, a hand-tracking device measures these postures, and corresponding virtual hand models will be displayed on the visualized simulated system. The virtual hand models behave in synchronization with the real hands, and the user can apply external force to atoms by touching them by the virtual hands.

Our system automatically switches multiple algorithms for calculating the external forces for each atom and operating simulation system, depending on the presenting hand gestures and the structures of the proteins, to provide the user with a highly intuitive experience of manipulating molecules with broad range of molecular sizes as if they are really touching objects floating on air.

For example, if a hand shows the “pinching” gesture, an atom or an atom group pinched by the fingers is pulled to the direction in which fingers move. A hand with “holding” gesture captures a secondary structure of a protein and steers them in 6-DOF, and if all fingers of a hand is stretched, all atoms touching the hand are given repulsion force from palm and fingers.

To verify the usefulness of the proposed method, we conducted a brief evaluation by user testing. The participants were given several molecular manipulation tasks such as protein unfolding, and measured the needed time to complete them using either MOLding or the default input method (mouse dragging).

The result of the evaluation suggested that the users were able to perform a simulation task using our method in comparable time with the conventional method, and in some cases, our method gives better performance. Using our system, the user might easily reproduce processes of protein folding and protein-small molecule interaction by their hands in simulated system for investigating mechanisms of biochemical phenomena.

J35: Sandie Arnoux, Yvon Jégou, Gaël Beaunée and Pauline Ezanno. A generic framework to model infection dynamics in a metapopulation of cattle herds

Abstract: Endemic infectious livestock diseases impact animal health and welfare, and food safety. Pathogens spread between farms mainly due to animal movements (purchases/sales) and neighboring relationships. The risk of spreading depends on the within-farm proportions of infected animals, which varies within and between farms over time. A modelling approach is relevant to represent such a complex biological system, permitting the ex-ante evaluation of control strategies under various scenarios. Developing epidemiological models at a regional scale requires to couple within-farm epidemiological models, leading to complex models and to the need for large computational resources, especially when stochastic processes are involved.

The objective is to find the best generic framework in terms of computational performance to represent pathogen spread in a cattle metapopulation. Three requirements should be fulfil: (1)

a common interface should be used to run population dynamics, and within- and between-herd infection dynamics; (2) a common data structure should be used for animal movements; (3) the shared interface and structure should be easy to understand, to be usable by persons with various skill levels in modelling.

Two implementations are available in this framework: synchronized or desynchronized methods. In the first case, herd dynamics evolve simultaneously. At each time step, dynamics are simulated for all the herds. The second case was conceived to be used with distributed computing. Herd dynamics evolve independently from each other as long as no purchase occurs. As a purchase corresponds to a sell, the destination herd has to wait until an animal is available and its infection status known. For the latter case, neighboring contacts cannot be modelled as herd dynamics are not synchronized.

These implementations have been tested on a single processor, then will be tested on a computing grid. We have investigated the computing load according to herd size, herd number, number of years, and distribution of animal movements. When on average one movement occurred per year and per herd, the desynchronized method was slower than the synchronized one. On the contrary, if only a few herds (1/5) exchange animals, for the same total number of movements, the desynchronized method was faster. We successfully applied the synchronized framework to Mycobacterium avium subsp. paratuberculosis, which is spread by animal movements. The next step is to evaluate it on a grid, before using it for other pathogens with other spread characteristics.

J36: Tor Johan Mikael Karlsson, Óscar Torreño Tirado and Oswaldo Trelles. jORCA: Jumping to the Cloud

Abstract: Background: Technological breakthroughs in biological and biomedical data acquisition, for example with Next Generation Sequencing (NGS) systems, are generating big amounts of data. Cloud Computing (CC)) lets researchers rent computational and storage resources (accessed as Web-Services (WS)) and is becoming a viable option for such data processing.

We report the development of several plugins for the jORCA client [1], based on MAPI [2], which can provide a flexible front-end for WS deployed on CC.

Methods: We have used two CC platforms:

- A community cloud installation for the Mr.SymBioMath project [3]. Data access is done via Globus Online (GO) [4] which is an efficient, fault-tolerant and secure data transfer method.
- The commercial Azure cloud [5] where robust data transfer is possible via a set of WS calls.

Results: Because of the big sizes involved, data should not be transferred as part of the WS call. Instead, we first upload the input data to the remote storage using a reliable protocol and, when ready, invoke the WS and start to process the data using the computational resources.

Data transfer extensions: The GO plugin requires a local installation of Globus Connect, which provides a local GridFTP server (representing an endpoint). Once the data transfer is initiated, GO manages all aspects and no further interaction is needed by jORCA.

In contrast, the Azure plugin directly controls the data transfers (i.e. jORCA interaction is needed). Authentication is made using a temporary keys which is sent to the WS (to allow it to retrieve the data). Data transfers are made using Azure data blobs which allows jORCA to upload parts of the data and, if necessary, resume interrupted transfers.

Call by reference to web-services: We have implemented a RESTful WS which is used in both platforms. The initial step is to submit the required parameters (data references to data already stored in the cloud). The submission creates a new resource (job) which can be polled for status and, when ready, for the final result (i.e. a new data reference). This greatly facilitates the invocation of a series of WS (i.e. a pipeline/workflow) by avoiding the need to download intermediate results.

References
1. Martin-Requena, V. et al. (2010). jORCA: easily integrating bioinformatics Web Services, Bioinformatics 26(4), pp. 553-559.
2. Ramirez, S. et al (2011). MAPI: towards the integrated exploitation of bioinformatics Web Services. BMC bioinformatics, 12(1), 419.
3. Mr.SymBioMath – http://www.mrsymbiomath.eu
4. Globus Online - https://www.globus.org
5. Azure - http://azure.microsoft.com

J37: Alexis Allot, Laetitia Poidevin, Kirsley Chennen, Raymond Ripp, Julie Thompson, Olivier Poch and Odile Lecompte. GeneBook: a social network linking genes, diseases and researchers

Abstract: With the constant and massive increase of biological information, efficient access to useful information and production of knowledge becomes a growing challenge for successful research processes (1). User-centric and collaborative approaches are emerging to extract information about genes, diseases or scientific literature with services like MIMmatch, GeneTalk, ClinGene, iHOP or PubTator.

Here, we present GeneBook, a social network that directly interconnects three types of actors: humans, genes and diseases. The goal of GeneBook is to optimize and speed up research processes by providing a user friendly web service for clinicians and biologists allowing retrieval, annotation and interaction with information related to genes, diseases and researchers. GeneBook deploys a human-friendly research hub aiming at integrating all necessary features in a single location. The platform built around the powerful Play! Framework uses SQL and graph databases (Neo4j) for information storage and processing, as well as recent web technologies such as SVG graphics, Bootstrap framework and Highcharts.js to ensure enhanced visualization.

The heterogeneous network: GeneBook allows users to become friends (publicly or privately) with humans, genes and diseases, manage this friendship, view new friendship suggestions, and follow the activity of their friends (a gene-friend befriends a new disease, a disease has it's personal information updated...). They can “like” their friend's activity, navigate the friend-of-friend graph and find humans, genes and diseases sharing same interests. Networking between genes and diseases is ensured by a daily data mining process through public databases and by human interactions.

Personalized context for information retrieval and knowledge extraction: The concept of “interest sessions” allows users to specify active research topics and organize genes, diseases, keywords and other information. Various analyses can be performed on such sets (like functional enrichment on genes). Most importantly, these data allow GeneBook to take into account user interests, and automatically adapt available structured or unstructured information such as publications or disease textual descriptions by filtering relevant parts, highlighting important data, suggesting research paths and producing meaningful knowledge.

Management of the research process: GeneBooks adapts mainstream management tools for efficient biological information organization and transfer. The user can annotate any gene or disease, create tasks as operators to link genes, diseases and humans, communicate through a messaging widget with recognition of bioinformatics service links, gene names, etc...

J38: Caroline Siegenthaler and Rudiyanto Gunawan. Assessing Inference Methods in the Absence of Gold Standard Networks: Can Crowdsourcing Help?

Abstract: The inference of biological networks is a highly active research area in systems biology. Many algorithms for network inference have been developed in the last decade,

underlining the importance of assessment and comparison among these methods. Meanwhile, biological network inference, particularly that of gene regulatory networks (GRNs), is typically underdetermined. The underdetermined nature implies that the inference problem does not have a unique solution and only a part of the GRN can be reconstructed from data. In this regard, the accuracy of network predictions depends not only on the ability of a method to infer direct gene regulations, but also on the availability of causal information in the data. However, the extent to which a GRN can be inferred from gene expression data has commonly not been considered in existing assessments.

We recently published a procedure for assessing inference methods that took into account the inferability of GRNs. In particular, we excluded gene regulations that were deemed non-inferable from the assessment. The inferability of GRNs was analysed by considering an ensemble of networks which were indistinguishable from the data. The inferability analysis provided the lower and upper bounds of the network ensemble, where the difference between the lower and upper bound networks corresponded to non-inferable gene regulations. We applied the assessment to the network predictions in the DREAM 4 in silico network inference challenge. For this challenge, we determined the inferability of the GRNs from steady-state expression data of single-gene knock-out experiments and constructed the ensemble bounds from the gold standard networks. Unfortunately, in practice the gold standard network is usually not available.

In this work, we modify our assessment procedure to accomodate the situation in which the gold standard network is not available. We employ a crowdsourcing strategy to determine the inferable part of the GRN. Community predictions obtained by averaging over the predictions from different methods have shown more robust performance, and can capture the true network structure better than any single inference method. Furthermore, grouping network predictions based on local topological similarities before performing the average or voting within each group, can also offer advantages. In the new assessment, we construct the lower bound of the network ensemble based on gene interactions with high confidence scores, and the upper bound based on (the complement of) interactions with low confidence scores. We explore and evaluate different community averaging or voting strategies to obtain the confidence scores from a set of network predictions. To illustrate the application of the new assessment, we re-evaluate the predictions of the DREAM 4 in silico network inference challenge using this assessment, and compare the results with the original and our recently published assessments.

J39: Felipe Albrecht, Christoph Bock and Thomas Lengauer. DeepBlue: Epigenomic Data Server

Abstract: High volumes of data for studying epigenetic regulation are being generated by epigenomic consortia, including ENCODE, Roadmap Epigenomics, Blueprint Epigenetics, and DEEP projects. New problems arise with this data deluge: how to store and distribute the data, how to handle the associated metadata, and how to perform different types of analysis on such data.

We developed the DeepBlue Epigenomic Data Server, an online Data Server for storing and working with genomic and epigenomic data, in order to help to address these questions. DeepBlue provides a means of storing, organizing, searching, and retrieving epigenetic data, addressing the following challenges: (i) coping with the expected increase in volume of available epigenetic data; (ii) keeping our in-house software EpiExplorer [1] and EpiGraph [2] up to date in a timely and efficient manner; (iii) making all data easily accessible in a standardized form to increase efficiency of epigenomic data analysis and software development. Among the DeepBlue features, we highlight: (i) preinstalled datasets from the major epigenomic consortia (ENCODE, Roadmap Epigenomic, and BLUEPRINT

Epigenomics), (ii) an update module that retrieves the latest epigenome datasets from the repositories of several epigenome consortia, (iii) support for analysis operations directly on the data server, (iii) reproducibility by automatically documenting and storing the analysis steps.

DeepBlue supports a set of analysis operations on the epigenomic data, and implements a controlled vocabulary to ensure the data consistency. The set of available operations includes: filtering epigenomic data by metadata and region attributes, finding overlapping regions sets, grouping regions, retrieving DNA sequences retrieval and pattern matching operations.

DeepBlue can be accessed via an XML-RPC protocol that is supported by all major programming languages. The data are stored using MongoDB [3]. We use MongoDB because it has a flexible data model and also provides scalability through data distribution across several computers nodes (sharding). We use DeepBlue as data server for our EpiExplorer and EpiGraph software, as well as internal epigenomic data storage and analysis tool. DeepBlue is available to external users upon request. The manual and API reference are available at http://deepblue.mpi-inf.mpg.de/. This work has been supported by German Science Ministry Grant No. 01KU1216A (DEEP project) and has been performed in the context of EU grant no. HEALTH-F5-2011-282510 (BLUEPRINT project)

1. Halachev, K., Bast, H., Albrecht, F., Lengauer, T. & Bock, C. EpiExplorer: live exploration and global analysis of large epigenomic datasets. Genome Biology 13, R96 (2012).
2. Bock, C., Halachev, K., Büch, J. & Lengauer, T. EpiGRAPH: user-friendly software for statistical analysis and prediction of (epi)genomic data. Genome Biology 10, R14 (2009).
3. MongoDB, Inc. MongoDB http://www.mongodb.org/

J40: Sarvesh Nikumbh and Nico Pfeifer. On the Hardness of Computationally Predicting Long-Range Chromatin Interactions

Abstract: It is well known that chromatin, a complex of DNA and proteins, is packed in 3D space inside the nucleus and that the spatial conformation of the chromosomes is non-random. Furthermore, it is closely correlated with the functional state of the cell and gene activity among other factors. Thus, a better understanding of this 3D landscape and the underlying mechanisms can help in gaining an enhanced comprehension of many genomic functions. Recent studies have shown a correlation between these long-range chromosomal interactions and the functional state of the cell (normal/diseased).

With the advent of 3C-based technologies in the last decade, more recently Hi-C, a genome-wide analysis of the interaction profiles is now possible. These interactions comprise pairs of loci that are close in the three-dimensional structure of the chromosome, but not necessarily in the DNA sequence. Still, very little is known about the folding principles of the chromosomes. The spatial co-localization of different chromosomal regions (cis as well as trans) can be due to a mix of factors viz. specific, direct contacts between two loci, nonspecific binding as a result of the packing of the chromatin fibre or co-localization due to functional association or having the same subnuclear structure.

In this work, we have made an attempt to computationally learn and understand these contact pairs from the underlying DNA sequences. By extracting sequence-based features using various string kernels, namely the weighted-degree kernel with shifts, the local alignment kernel, the oligo kernel and the oligomer-distance histograms kernel, we aim to learn to classify the given set of linearly distal regions into spatially proximal or distal to a particular TSS-containing region. As a result of the aforementioned factors and the resolution of the Hi-C experiments, in having to uncover the potential true causal sections of the interacting chromosomal region pairs, also known as topologically associated domains (TADs), we have to deal with issues of 1) large variation in the length of the reported contact regions 2) distributional characteristics of the multiple candidate regions, and 3) intelligently weeding out the noisy portions. While (1) rendered some of the available kernels, e.g. weighted-degree with shifts kernel, useless, as they require the two sequences being compared to be of the same length, (2) affected all kernels. We devised an approach that could help us tackle these and (3) by focusing on the candidate portions of the complete reported chromosomal regions from the experiment. We present on 1) the results of this early work, 2) the hardness of computationally predicting these long-range interactions, and 3) on the next steps of why one would require supplementary information like histone modifications, chromatin states etc. and how one can supplement the sequence-based features with these to possibly improve the performance of such a system.

J41: Kazuki Kishi, Daisuke Komura, Takayuki Isagawa and Shumpei Ishikawa. Visualizing whole cancer-stromal interactome

Abstract: Cancer cells usually survive in microenvironment surrounded by non-cancer “stromal” cells. Cancer-stroma cell interaction is suggested to be significant for cancer survival and progression. Although there has been many reports about molecules responsible for the cancer-stroma interactions, lack of methods for systematic and quantitative profiling of the whole interactome makes it difficult to compare different interactions with each other and prioritize a particular interaction for subsequent clinical approach.

Through transcriptome sequencing of cancer xenograft tissues, we differentially assign human/cancer cell- and mouse/stromal cell- derived transcripts, integrate public protein-protein interaction database, and construct a cancer-stroma interactome map. By comparing the mapping scores of the paired-end read, we got reliable differentiation between human- and mouse- homologous transcript. Furthermore, to get expression and interactome profile robust for tissue degeneration and experimental conditions, we developed methods for modifing raw mapping count using several parameters like GC contents, length from poly-A tails and originally defined “mappable regions among annotated transcripts”. This map shows us with quantitative information about how a particular interaction contributes to the whole interactome.

For the purpose of visualizing the interactome map easily and interactively, we implemented a system where we can search a particular interaction, set the threshold for visualization and analyze the data in a single window. There are 4 visualization canvases in a window, each of which shows different output. In a canvas, each interaction is visualized as bubbles which position and radius means the direction and intensity of the interaction, respectively. This system is implemented as a web application, with capability of simultaneous access.

We applied this tool for our cancer xenograft data sets. These data shows diverse cancer-stroma interactome profile among different cancer types, and even single particular interaction has variable contributions among the whole interactome. While the well-known interactions, which are targets for established molecular-drugs, shows strong interactions between cancer and stroma cells, we found several other interactions have potentially stronger contributions, which could be new therapeutic target.

These data suggest that whole cancer-stroma interactome visualization is useful approach for discovery of new cancer targets.

J42: Nicolas Tchitchek and Christophe Becavin. MDS-Reference Maps and MDS-Voronoi Representations for Visualization and Analysis of High Dimensional –Omics Profiles

Abstract: High-throughput expression data from –omics experiments are complex to analyze because of the large number of biological features measured. For each –omics profile, thousands of different variables are measured at a single time, such as with transcriptomics profiles where up to 40,000 gene expression values are measured. Dimensionality reduction

methods, such as Multidimensional Scaling (MDS), allow to project –omics profiles in 2 dimensional spaces to visualize similarities and distinctness between the samples. In MDS representations, each dot is the –omics profile of a biological sample and pairwise distances between dots are proportional to the biological distances between the samples.

We present here two extensions for MDS methods that allow (i) to visualize profiles relative to reference datasets for meta-study analysis; and (ii) to represent features driving the similarities and distinctness between the profiles.

The first extension allows the projection of new –omics profiles over a predefined MDS representation (named MDS Reference Map). The resulting representation (named MDS projection) allows then to visualize the similarities and distinctness between samples in regards to another study, with the advantage of having consistent reference representations with different analyses. MDS Reference Maps and MDS Projections have been generated using a molecular dynamics based algorithm. In a first step, the method performs a dimensional reduction of objects by modeling objects by particles and pairwise distances between them by repulsion and attraction forces. In a second step, the method assigns an infinite mass to each object (i.e. particle) of the MDS Reference Map, resulting in a projection of the additional objects over the reference representation.

The second extension allows to overlay expression data into MDS representations. The resulting representations are named MDS-Voronoi representations. MDS-Voronoi representations take advantage of natural space division obtained with Voronoi diagrams. In a Voronoi diagram, a set of points is specified and a region, called a Voronoi cell, is defined for each point. This region consists of all points closer to that point than to any other. In our approach the set of points corresponds to the MDS representation of a set of –omics profile. The MDS-Voronoi cells are then overlaid with expression values to visualize what drives the similarities and distinctness between biological samples.

We show the relevance of these two new approaches with three publicly available transcriptomics datasets of: (i) thirty-two profiles from different human tissues; (ii) seventy-two profiles of Mouse Embryonic Fibroblasts cells obtained from different genomes; and (iii) three hundreds profiles of mouse lungs infected by different influenza viruses.

J43: Frédéric Mahé, Torbjørn Rognes, Christopher Quince, Colomban de Vargas and Micah Dunthorn. Swarm: robust and fast clustering method for amplicon-based studies

Abstract: High-throughput sequencing technologies are today our best approach to deeply assess the environmental or clinical diversity of complex assemblages of microorganisms. Because of the increasing sizes of amplicon datasets, fast and greedy de novo clustering heuristics are the preferred and the only practical approach to produce molecular operational taxonomic units.

These clustering methods suffer from two fundamental flaws: arbitrary global clustering thresholds, and input-order dependency induced by centroid selection. Swarm was developed to address these issues by first clustering nearly identical amplicons iteratively using a small local threshold, and then by using clusters' internal structure and amplicon abundances to refine its results. This fast, scalable, and input-order independent approach reduces the influence of clustering parameters and produces more robust operational taxonomic units than other clustering methods. Results obtained on large scale environmental studies (e.g. TARA OCEANS, Neotropical Forest Soils) will illustrate the interesting properties of swarm, and the new insight it gives us on amplicon data.

J44: Christofer Bäcklin and Mats Gustafsson. Developer friendly and computationally efficient predictive modeling without information leakage: The emil package for R

Abstract: Machine learning-based solutions to predictive modeling problems in biology and medicine (classification, regression, survival analysis) typically involve a number of steps beginning with data pre-processing and ending with performance evaluation. A large number of packages providing tools for the individual steps are available for R but not for facilitating the assembly of them into complete modeling procedures or rigorously evaluating their combined performance.

We present a new package for R denoted emil (evaluation of modeling without information leakage) that is designed to be a flexible backbone of modeling procedures having the following properties: (1) Enable evaluation of performance and variable importance by means of resampling methods without introducing information leakage. (2) Return parameter tuning statistics and final prediction models. (3) Transparent, highly customizable and easy to debug structure. (4) Offer the user direct control over memory and CPU-intensive steps of the calculations. (5) Comprehensive yet concise documentation.

To ensure practical usability of the package despite the complex nature of the task it addresses, it was developed according to two design principles. Firstly, the user should be able to quickly trace errors that occurs during the modeling and debug the functions that caused them, even without in-depth knowledge of the package. Secondly, the user should be empowered to re-implement any aspect of the package for which there exist no universally superior implementation. Examples of such is parallelization, resampling method, or parameter tuning routine. The result is a computationally efficient and developer friendly framework that enables resampling based analyzes using millions of variables, is easy to extend, and allows development of scalable solutions.

We explain emil's functionality in the context of standard usage, resampling, and customization. Specific application examples are presented to show its potential in terms of parallelization, customization for survival analysis, and memory management.

J46: Guoxian Yu, Hailong Zhu and Carlotta Domeniconi. Predicting Protein Functions using Incomplete Hierarchical Labels

Abstract: Predicting protein functions is a hard problem, characterized by several factors: (1) the number of function labels is typically large; (2) a protein may be associated with multiple labels; (3) the function labels are structured in a hierarchy; and (4) the labels are incomplete. Current predictive models often assume that the labels of the labeled proteins are complete, i.e. no label is missing. But in real scenarios, we may be aware of only some hierarchical labels of a protein, and we may not know whether additional ones are actually present. The scenario of incomplete hierarchical labels, a challenging and practical problem, is seldom studied in protein function prediction (Valentini, 2014; Yu et al., 2014).

In this poster, we propose an approach to Predict protein functions using Incomplete hierarchical LabeLs (PILL in short). PILL takes into account the hierarchical and the flat taxonomy similarity between function labels, and defines a Combined Similarity (ComSim) to measure the correlation between labels. PILL estimates the missing labels for a protein based on ComSim and the known labels of the protein, and uses a regularization to exploit the interactions between proteins for function prediction. PILL is shown to outperform other related techniques in replenishing the missing labels and in predicting the functions of completely unlabeled proteins on publicly available PPI datasets labeled with Gene Ontology (Ashburner et al., 2000) and MIPS FunCat labels (Reupp et al., 2004). In addition, our real life example shows that for 451 recently (from 2014-02-01 to 2014-06-01) appended GO terms associations (proteins and GO labels associations) of S. Cerevisiae, PILL can correctly replenish 141 labels and these labels have been supported by PubMed articles. The datasets and codes used for PILL are available upon request.

References

[1]Valentini G. (2014) Hierarchical ensemble methods for protein function prediction. ISRN Bioinformatics, Vol. 2014, Article ID 901419, 34 pages, 2014. doi:10.1155/2014/901419
[2]Yu et al. (2014) Protein function prediction using incomplete annotations. IEEE/ACM Trans. Computational Biology and Bioinformatics, 99(1), 1-1.
[3]Ashburner M. et al. (2000) Gene Ontology: tool for the unification of biology. Nature Genetics, 25(1), 25-29.
[4]Ruepp A. et al. (2004) The FunCat, a functional annotation scheme for systematic classification of proteins from whole genomes. Nucleic Acids Research, 32(18), 5539-5545.

J47: Jan Grau, Jens Boch and Stefan Posch. Genome-wide TALEN off-target prediction and its utility for TALEN design

Abstract: Transcription activator-like effector nucleases (TALENs) have become an accepted tool for targeted mutagenesis. The DNA binding domain of TALENs originates from transcription activator-like effectors (TALEs) of plant-pathogenic Xanthomonas bacteria. TALEs are injected by Xanthomonas into host plant cells to function as transcriptional activators for the benefit of the pathogen. The DNA-binding domain of TALEs and thus TALENs is composed of conserved amino acid repeats containing repeat-variable diresidues (RVDs) that determine DNA binding specificity.

In TALENs, this DNA-binding domain is fused with a Fok1 endonuclease domain, where homo- or hetero-dimers of TALENs specifically cut the DNA double strand.

Although TALENs cut DNA highly specific, undesired off-targets in addition to the targeted genomic region remain an important issue, since these may cause severe side effects due to off-target mutagenesis.

We developed a tool for the genome-wide prediction of TALEN off-targets, named TALENoffer. TALENoffer is based on the same statistical model as TALgetter, a program that successfully predicts TALE targets in host promoteromes. In TALENoffer, this model is employed to score TALEN monomers. Monomer scores are combined into a dimer score that compensates for differing monomer lengths and that allows for partial compensation between monomers for target site mismatches. In benchmark studies, TALENoffer yields a competitive performance compared to alternative approaches.

TALENoffer provides flexible options to adjust the scanning process to specific TALEN architectures as for instance TALENs with obligatory hetero-dimerization or non-standard TALENs where the Fok1 domain is fused to the N-terminus of the DNA-binding domain. Users may also specify the range of distances between TALEN monomers that yield functional dimers with presets for several common TALEN architectures.

TALENoffer features strategies for runtime optimization, which allow to scan complete genomes for TALEN off-target sites within a few minutes on a standard PC.

We make TALENoffer available as a web-application at http://galaxy.informatik.uni-halle.de, which can also be installed in a local Galaxy server, e.g., for confidential analyses. In addition, we provide a command line version of TALENoffer at http://jstacs.de/index.php/TALENoffer, which is easily scriptable for high-throughput analyses.

The off-target predictions of TALENoffer are one component of TALENdesigner, a novel tool for designing highly active TALENs that target desired genomic loci with minimized off-target effects. TALENdesigner is currently under development and we give an overview of its features and performance.

J48: Iris Leitner and Oswaldo Trelles. Intuitive library for efficient access of compressed genome sequences

Abstract: Background: The amount of collected data is constantly growing in all areas of science and with it the need for compression. Several algorithms and tools for sequence compression have therefore been developed. In Bioinformatics there are a number of tools [2] implementing compression methods specifically for single FASTA files. While these tools all have different strength and weaknesses, most of them focus on providing compression and decompression on full files and do not allow processing the compressed sequences directly, which could be a big advantage, especially for the work with genome sized sequences.

Methods: We have developed an application which combines compression and decompression of FASTA files (libracio) with an intuitively usable C-stdio like interface (SFILE) with the ability to access data directly from the compressed sequence without the need of decompressing the full file.

Libracio (Random Access I/O Library) implements a FASTA specific encoding strategy and combines it with standard compression methods. To provide the ability of random access, a block-wise compression strategy was implemented. We parallelized compression, decompression and random access in order to take advantage of modern multiprocessors.

The SFILE library implements the classical I/O methods available in file systems, such as open, close, read, seek, etc. and uses Libracio as the backend for accessing compressed FASTA sequences. Additionally, it provides sequence information and implements a caching system to speed-up likely access patterns.

Results: Initial benchmarking tests show that even without the use of FASTA specific encoding strategy, the compression and random access time is significantly improved in comparison to standard gzip [1] compression, although with a slight loss in the compression rate.

Our statistical analysis of the data set presented in [2] suggests that with encoding we can reduce more than 90% of all ‘N’s in the sequences to a few bytes. For the rest of the sequence we expect a reduction of 50% to 75% with depending on the sequence structure. By applying an additional compression method to the encoded sequence we expect an improvement in the compression rate and random access time.

Conclusion: Libracio and SFILE create in combination a well-tested application, which provides reliable compression and decompression of FASTA files and an intuitive API to provide I/O functionality. The biggest advantage lies in the ability to directly process compressed sequences, without the need of decompressing the full file, which is especially useful when working with genome sized sequences.

References

[1] http://www.gzip.org/

[2] Zhu Z., et al (2013). High-throughput DNA sequence data compression Briefings in Bioinformatics

J49: Jose Arjona-Medina, Óscar Torreño Tirado and Oswaldo Trelles. Software for featuring genome evolution

Abstract: Motivation: The post-genomic era is characterized by the release of large amounts of data boosted by the scientific revolution in high-throughput technologies. The increase in the availability of complete sequenced organisms, and the importance of evolutionary processes affecting the species history, has increased the interest in studying molecular evolution events (EE).

These EE, also named ‘genome rearrangements’, play a crucial role in evolution, particularly in bacteria. From a computational point of view, the characterization of EE (frequencies, average length, distance between origin and current position, etc.) is a challenging work since it involves different complex computational steps. Addressing this problem, we can obtain potential benefits in the estimation of genome distances, the construction of evolutionary

trees, and phenotype-genotype correlation studies, amongst others.
Methods: The methodology we envisage begins with the fast detection of High-scoring Segment Pairs (HSP). Since small evolutionary events such as indels and mutations can produce different but close HSPs, a second step is aimed at defining what we name Computational Synteny Blocks (CSB) –that groups close HSPs. In a third step we identify the type of event, at this moment concentrated in Translocations (T), Inversions (I) and Inverted translocations (IT). In a final fourth step we estimate the main features for each type of EE.
Results: We have benchmarked our proposal with a set of 66 mycoplasma bacteria, performing 2.145 full genome comparisons. The process starts with the genome sequences, and ends with the characterization of EE which takes 16 seconds.
Our initial results show that the PDF (probability density function) of CSB length undergoing EE is far from random.
Conclusions: We have designed specific and efficient software that is not constrained by sequence length. EE frequencies are efficiently measured without gene permutation information, in contrast with most state-of-the-art software (GRIMM (Tesler, G., 2002), Cinteny (Sinha, A. et al. 2007)) and without any parameters. This could be used in several biological applications including estimation of divergence, improved phylogeny models, etc.
Evolution modifies genome organization using events such as Translocations, Inversions and Inverted translocations, amongst others. The success of the rearrangements does not follow a random basis; some organizations have a higher probability than others, as suggested by our estimated PDFs.
Although this is an initial study focused on a specific dataset, our results intuitively match with expected results. The obtained results could be strongly associated with the dataset; therefore we are working on a more exhaustive study to validate results.
This study and resulting conclusions are based on a medium size dataset, and we are reasonably confident that, at least in outline, the picture that emerges is not too far from the truth.

J50: Hanna Ćwiek, Augustyn Markiewicz and Paweł Krajewski. Phenalyser: a web-based ISA-TAB-compliant tool for analysis of phenotyping experiments

Abstract: Phenotype analysis is an important aspect of plant research. In combination with genetic information, phenotypes are basis to discover the nature of the processes taking place in plant organisms. Phenotype data collection is an expensive and time consuming process, as growing plants in experiments takes a lot of time and space, requires precise conditions, and observations and measurements are meticulous. Huge effort put in phenotyping experiments deserves scrupulous handling with the results, so that they can be used in multiple analyses, both now and in the future, together with the fast growing set of genetic data (which, contrary to phenotypes, seem to be well curated, ordered and synchronised within a number of resources collaborating across the world). Proper annotation and comprehensive formatting is necessary to exchange, understand and analyse one another's experiments' results.
We present Phenalyser - a web application for statistical analysis of phenotypic data. Our tool uses a unified format for description of phenotypic experiments - an extension of ISA-TAB format for phenotyping. The analysis of phenotypes is assisted by ontological description of the data provided in the datasets. Currently, the tool computes sufficient statistics for factorial phenotyping experiments based on evaluation of linear mixed models for phenotypic traits. The results, consistently formatted and thus comprehensive, enrich the datasets. They can be used in standard analysis of plant experiments and serve as a means to reduce data size for further storage. The application also serves as a short-term storage of phenotypic data, thanks to the adaptation of BII database modules provided by ISA-tools.
This research has received funding from the European Union's Seventh Framework

Programme for research, technological development and demonstration under grant agreement no 283496, project transPLANT (http://www.transplantdb.eu), coordinated by EMBL-EBI.

J51: Helen Lindsay, Alexa Burger, Jonas Zaugg, Christian Mosimann and Mark Robinson. An R toolkit for studying the CRISPR-Cas9 mutation spectrum

Abstract: The CRISPR-Cas9 system is an efficient method of introducing targeted mutations into genomic DNA sequences and has been recently applied to a variety of species including human, zebrafish, and yeast. A single-stranded guide RNA (sgRNA) directs the Cas9 nuclease activity to a 20 nucleotide target region. Repair of cleaved DNA sequences by non-homologous end-joining introduces a variety of variants, including insertions and deletions, while co-delivery of a repair template can be harnessed towards precise genome editing. Existing multiple sequence alignment tools are not ideally suited for visualising a collection of pairwise alignments of variant sequences to a reference genome, as conflicting insertion sequences can occur. Particularly as high-throughput methods of CRISPR-Cas9-mediated mutagenesis are developed, new methods of analyzing this mutagenesis spectrum will be needed to validate mutagenesis efficiency and variants. We have developed an R-based toolkit for counting and visualising variants of a target region. Our software takes a set of BAM, FASTQ, or AB1 format files routinely used to sequence target loci, and outputs summary metrics and plots of variant combination frequencies, as well as a custom reference-based multiple sequence alignment of the region surrounding the target. This output allows a compact representation of the potentially large number of variants. With this toolkit, we aim to assist researchers in characterising the efficiency of their mutagenesis system using standard and high-throughput sequencing methods, assessing off-target effects, and ultimately understanding how the mutations are induced.

J52: Maciej Pajak, Clive Bramham and Ian Simpson. Computational approaches to improving miRNA-mRNA interaction predictions

Abstract: MicroRNAs (miRNAs) are 20-22nt long transcripts that form protein complexes that bind to mRNA and decrease the translation rate of the target mRNA. An ongoing problem in miRNA research is target prediction, as identifying targets is necessary for biological interpretation of the findings. Multiple computational algorithms exist, however, their usefulness is debatable. Although most of the competing methods use the same basic principles their outputs have very poor convergence. Recently, an alternative method of miRNA target identification based on deep sequencing of crosslinked samples and identification of chimeric reads (CLASH) was successfully applied to a subset of human miRNAs. This study demonstrated that there are several other classes of miRNA-mRNA binding sequence (outside the widely cited seed region) and revealed an abundance of non-canonical sequences (Helwak, et al., 2013). The emergence of these novel signatures for miRNA binding sites creates the opportunity to develop novel prediction methods augmented by CLASH datasets.

We present ‘targetPredictor’, an R package implementing an aggregation method for miRNA target prediction data that results in improved predictions and reduces noise when compared to data from any single miRNA target prediction source. We assessed integration methods using precision-recall curves against ‘gold-standard’ binding data from miRTarBase. The ‘targetPredictor’ package facilitates the implementation of workflows requiring miRNA target prediction through a supplementary annotation package containing human and mouse target predictions and for other species through homology transfer when direct predictions are not available from MiRanda, TargetScan, DIANA or PicTar. Finally, we propose an alternative

approach to miRNA target prediction, isolating sequence signatures from the direct miRNA-mRNA binding data of CLASH experiments and training classifiers to reveal likely binding sites. Despite the current paucity of high quality CLASH data upon which to train, these data are accumulating rapidly in multiple species and conditions and are likely to greatly improve the accuracy and reliability of miRNA target prediction in the near future.

J53: Chen Meng, Bernhard Küster, Aedín Culhane and Amin Moghaddas Gholami. Integration of multiple omics data for detecting cluster specific perturbed gene sets

Abstract: Nowadays, increasing number of studies, including The Cancer Genome Atlas (TCGA) and The Encyclopedia of DNA Elements (ENCODE), profile multiple levels of biological molecules from a large number of samples. One of the important goals of analyzing these data is to identify meaningful clusters. However, the large number of variables in omics data obstructs interpreting biological relevance of clusters. Hence, gene set analysis (GSA) is usually applied as its greater biological interpretability. Some unsupervised gene set based approaches were proposed to analyze large dataset without relaying on replicates, such as gene set variation analysis (GSVA, Hänzelmann et al. 2013) and single sample gene set enrichment analysis (ssGSEA). Nevertheless, all these methods do not benefit from the potential of integrating multi-omics data. To address this challenge, we introduce Multi-Omics Gene Set Analysis (MOGSA), a new method takes advantage of integrating multiple omics data and gene set annotation to facilitate clustering discovery and identify cluster driven gene sets.

MOGSA algorithm consists of three steps. First, multiple data are integrated via a multi-table method, such as multiple factorial analysis (MFA) or STATIS. Then, gene set information are projected as supplementary tables and meaningful principal components are used to reconstruct a gene set-sample matrix. Finally, we employed the truncated total bootstrap procedure (Cadoret & Husson, 2012) to estimate the significance level of a gene set in different samples.

We evaluated the performances of the method on both simulated and real biological data. Using simulated data, we compared MOGSA with existing methods and showed that 1) MOGSA outperforms naïve matrix multiplication because it is more tolerant to noise in data; 2) both specificity and sensitivity of detecting perturbed pathway are increased through integrating multiple datasets. In a real biological data study, MOGSA integrated microarray and proteomic data of NCI-60 cell lines and suggested that cell lines are segregated according to their tissue of origins on the gene set level. More importantly, it identified novel pathways that are missed by separate gene set analysis. The second real data study is the joint analysis of TCGA Bladder Cancer (BLCA) omics data, namely, DNA methylation, copy number variation (CNV), mRNA and protein data. In combining with consensus clustering, we distinguished 4 robust subtypes in BLCA, two of which resemble the proliferative and immunoreactive subtypes that were defined in ovarian cancer, suggesting a similar mechanism of development between the two heterogeneous cancers. In addition, the miRNA and transcriptional factor (TF) annotation of subtypes confirmed the importance of miR-99a, -100, -145, -125b and suggested novel subtype miRNA and TF markers. An R package implementing the algorithm is available through bioconductor.

J54: Jorge Álvarez-Jarreta and Gregorio de Miguel Casado. PhyloViewer: A Viewer for Large Phylogenies

Abstract: Phylogenetic trees have being widely used in nearly every branch of biology for modelling and analysing evolution over time (1; 2). In this field, there exist some reliable tools for inferring large phylogenies from big input data sets. This is the case of ZARAMIT

(3), for reconstructing the human mitochondrial phylogeny from raw DNA sequences or more generic systems such as SATé (4) and DACTAL (5).

However, real life exploitation of the phylogenies obtained is limited by the lack of specific tools able to provide the biologists with enhanced navigation and search features. To our knowledge, there are no tools available for phylogenies >1K nodes and also link additional biological data (structural features, metadata or fine grain details).

The tool PhyloViewer is presented as a visual add-on of PhyloFlow (6). It consists of a client/server architecture based on a web interface and a database server. A set of requirements provided by a group of biologists working with mtDNA and a detailed analysis of the technical features required to deal with trees >50K nodes has being considered. The system tests have being done with both synthetic and real data (a 8K node tree from ZARAMIT). Relevant features:

1. Interactive navigation over a main panel, which provides a father-siblings view.
2. Direct sequence search and historic node view from a working session.
3. Grouped tree view (by haplogroups for the case of human mtDNA).
4. Complementary visualization of specific node data (extended information from GenBank) and user annotations.
5. User login with profiles keeping private and public view phylogenies.
6. Backup and direct edition of phylogenies.
7. Administrative tools for the database and user management.

This tool would be available by the end of July 2014 in http://zaramit.org/.

Acknowledgement: The authors thank Samuel Guajardo, who contributed to the implementation, and Eduardo Ruiz Pesini, from the Mitochondrial Biogenesis and Pathology research group of the University of Zaragoza.

Funding: Project TIN2011-27479-C04-01 and grant FPU AP2010-1058.

References

[1]E.S. Allman and J.A. Rhodes, “Trees, Fast and Accurate.” Science, vol. 327(5971), pp. 1334–1335, 2010.

[2]J. Felsenstein, “Inferring Phylogenies.” Sinauer, 2003.

[3]R. Blanco, and E. Mayordomo, “ZARAMIT: a system for the evolutionary study of human mtDNA.” LNCS, vol. 5518(2), pp. 1139–1142, 2009.

[4]K. Liu, S. Raghavan, S. Nelesen, C. Linder, and T. Warnow, “Rapid and Accurate Large-Scale Coestimation of Sequence Alignments and Phylogenetic Trees.” Science, vol. 324, pp. 1561, 2009.

[5]S. Nelesen, K. Liu, L-S. Wang, C. Linder and T. Warnow, “DACTAL: divide-and-conquer trees (almost) without alignments.” Bioinformatics, vol. 28(12), pp. i274–i282, 2012.

[6]J. Álvarez-Jarreta, G. de Miguel Casado, E. Mayordomo, “PhyloFlow: A Fully Customizable and Automatic Workflow for Phylogeny Estimation.” Submitted to ECCB2014 poster session.

J55: Maarja Lepamets, Lauris Kaplinski, Reidar Andreson and Maido Remm. GenomeTester 4.0 – a k-mer analyzer package for sequencing reads and genomes

Abstract: We have developed a new unique k-mer based software package GenomeTester 4.0 for the analysis of genomic and sequencing data for various applications.

GenomeTester 4.0 package contains 3 separate programs:

* glistmaker parses sequence data in FastA or FastQ format and builds k-mer frequency table.
* glistquery searches k-mers from frequency tables.
* glistcompare performs set-theoretic operation on k-mer frequency tables

The counting of k-mers is implemented by sorting and collating k-mer list with O(n) complexity. In memory-tight situtations the maximum temporary table size can be capped

with the slight penalty to the calculation time. The maximum allowed word-length is 32.
The unique feature of GenomeTester 4.0 is the ability to perform set-theoretic operations of union, intersection and complement on k-mer frequency tables. By combining these elementary operations one can implement more complex algorithms and generate specific k-mer sets. This allows Genometester to be used as the basis of various other genomic applications - for example finding repeats, finding unique tag sequences and identifying species from sequencing reads.

For certain operations GenomeTester can use comparison with specified number of mismatches instead of identity. This makes it possible, for example, to find unique tag sequences for certain species that differ at least by at specified number of positions from the control group.

We have evaluated GenomeTester with both bacterial and human genomes and bacterial sequencing reads. It's performance while counting k-mers is comparable with other available tools. For example counting all 32-mers of full human reference genome takes around 20 minutes. Finding the set of unique 32-mers in E.coli genome compared to all other bacterial genomes from NCBI database took less than 9 minutes. Finding unique tag k-mers from 100GB bacterial sequencing reads took less than 10 minutes.

GenomeTester 4.0 will be available under free license from University of Tartu Bioinformatics group.

J56: Kerstin Johnsson, Jonas Wallin and Magnus Fontes. Model-based mutual clustering of flow cytometry data through Bayesian hierarchical modeling

Abstract: Flow cytometry is a widely used technology in which characteristics of each individual cell in a sample are measured. It has a plethora of applications within biomedicine, among other things it has been successful at elucidating the workings of the immune system and it is used clinically to monitor HIV progression and diagnose leukemia patients. The number of parameters that can be measured has grown steadily and it is now possible to quantify 18 different markers simultaneously using fluorescence. With mass cytometry, sometimes called next generation flow cytometry, it is already possible to quantify over 30 different markers for each cell.

As the dimensionality of the data sets increases the importance of automated data analysis methods rise. A main step in the analysis, which is still most often done manually, is to identify the relevant cell populations and determine the sizes of these. A key challenge that distinguishes this from the traditional clustering problem is to align corresponding cell populations between different samples. There are three different approaches to do this alignment. The first approach is to pool data from all the samples and perform clustering on this set. This has the disadvantage that clusters can be hard to distinguish in the pooled data since the clusters can be shifted between samples, it can also lead to complicated cluster shapes even when the individual cluster shapes are simple. The second approach is to cluster all samples individually and then match the clusters between samples. This means that no information is shared between corresponding clusters from different samples, furthermore if one or more clusters is absent in a sample severe complications arise. The third approach is to use a Bayesian hierarchical model for the clustering so that information between clusters in different samples are shared. Previously this has been done in a way where the cluster centers and the covariance matrices are fixed between samples. This means that it cannot take into account that clusters might be slightly shifted between samples and have slightly different shapes, which means that some of the problems with pooling the data also applies for this approach. We propose a new Bayesian hierarchical model where the cluster centers and covariance matrices are connected between samples in the model formulation. The inference of the model is taken from the posterior distribution, which is generated by a Gibbs sampler.

The model is fitted on a real world data set from the R package healthyFlowData. Our method detects the five relevant cell populations in the data set and we see that the model is appropriate. The inference for the model is well suited for parallelization and can thus be scaled to handle the large data sets that are common in flow cytometry.

J57: Damien Correia, Olivia Doppelt-Azeroual, Jean-Baptiste Denis, Mathias Vandenbogaert and Valérie Caro. MetaGenSense: A Web application for analysis and visualization of High throughput Sequencing metagenomic data

Abstract: The detection and characterization of emerging infectious agents has been a continuing public health concern. High throughput sequencing (or Next-Generation Sequencing; NGS) technologies have proven promising approaches for the unbiased detection of pathogens in complex biological samples. They are efficient and provide comprehensive analyses. However, NGS yields millions of putatively representative reads per sample, such that efficient data management and visualization resources have become a mandatory requirement, through a dedicated Laboratory Information Management System (LIMS), solely to provide perspective regarding the information contained in this huge amount of data.

We developed a managing and analytical bioinformatics framework that is engineered to run associated and dedicated Galaxy [1] workflows for the detection and eventually classification of pathogens. In essence, our primary purpose is to assist the biologist in the process of deciding on the most relevant sample-specific sequences in the supplied samples, and determine their relative abundance. To this end, a user-friendly interface is essential. A complete set of specific Galaxy pipelines, producing high quality reads and/or assemblies meaningful for biological interpretation, have been engineered, and serve as the driving engine for a graphical web-interface associating the sample’s meta-data and its analysis results. This user-interface has been tailored to associate a bio-IT provider’s resources (a Galaxy instance, sufficient storage and grid computing power), with the input data and its metadata. Hence, the web application allows scientists to easily interact with existing Galaxy metagenomic workflows, facilitates the organization, visualization and aggregation of the most significant and most meaningful bits of information from millions of genomic sequences. In more detail, communication between our Django-based interface [2] and Galaxy uses the Bioblend API [3], which gives access to a Galaxy instance’s main features, through scripted and automated commands. Metadata about samples, runs as well as the workflow results are stored in the LIMS.

Visualization tools associating the sequencing raw data with the analysis results are as important as the analysis itself. The interface already integrates existing tools such as KRONA [4], and also enables sharing of scientific results with several project members. In the end, it will also allow the integration of new visualization tools (in development).

[1] Goecks J, Nekrutenko A, Taylor J and The Galaxy Team. Genome Biol. 2010 Aug 25;11(8):R86.

[2] Django core team. URL http://www.djangoproject.com, 2011.

[3] C. Sloggett, N. Goonasekera, and E. Afgan, Bioinformatics (2013) 29 (13): 1685-1686.

[4] B.D. Ondov, N.H. Bergman and A.M. Phillippy, BMC Bioinformatics 2011, 12:385.

J58: Nadia El-Mabrouk, Laurent Gueguen, Manuel Lafond, Emmanuel Noutahi, Jonathan Séguin, Magali Semeria and Eric Tannier. Genome-wide gene tree correction

Abstract: Gene trees represent the evolutionary relationships between sets of homologous genes. They are useful to unveil the molecular evolutionary events that have shaped today's genomes. Although inferring phylogenies is a field with a very long history, due to various limitations, constructing good gene trees is still challenging. We propose a general methodology for gene tree correction based on genomic constraints, that can be applied subsequently, to gene trees output from any given reconstruction software. Potential errors in gene trees are detected from weakly supported edges, nodes identified as dubious for incongruence with the species tree or contradictions between the orthology/paralogy relations inferred from the gene tree and the gene order context. The correction is based on extensions of various tools developed in our labs, such as "PolytomySolver" which finds the best refinement of a polytomy minimizing the reconciliation cost, and "ParalogySolver" which computes a gene tree as close as possible to the initial one, satisfying orthology constraints. Corrected trees are then evaluated according to maximum likelihood computed from gene sequence alignment, and to the shape of reconstructed ancestral genomes. Results given for the "Ensembl Compara" gene trees clearly show an improvement of the trees following our correction methodology, according to various evaluation criteria.

J59: Eugenio Mattei, Fabrizio Ferrè and Manuela Helmer-Citterich. BEAR-Suite: a collection of tools for RNA structural comparison

Abstract: Motivation: Comparing RNAs using only the nucleic acid sequences can be inaccurate, especially when the sequence identity is <50-60%[Gardner et al.(2005)].Improving performances requires including secondary structure information during the computation. Affix arrays[Meyer et al.(2011)],tree-based representations[Blin et al.(2007)],simultaneous folding and aligning of RNAs[Will et al.(2007)],covariance models[Nawrocki et al.(2009)] are examples of approaches used to integrate sequence and structure information.These methods suffer from high computational and time resources, or setting up problems.Recently,we developed a new context-aware structural encoding for RNA secondary structure,called BEAR (Brand nEw Alphabet for RNA)[Mattei et al.(2014)].This encoding allows representing secondary structure information using an expanded alphabet that is able to discriminate among different RNA substructures. As such, RNA secondary structure can be described by an informative string of characters.Moreover,starting from the BEAR representation of related RNAs extracted from the Rfam[Gardner et al.(2011)] database,we captured transition rates between substructures and expressed them in the form of a substitution matrix(denoted as MBR) for secondary structure elements.
Methods: Here we present a suite of programs to perform multiple and pairwise alignment of RNAs exploiting the BEAR encoding and the MBR.The ability to encode secondary structure elements using a string of characters combined with the substitution matrix allows us to compare directly two RNAs exploiting the same algorithms used in nucleic acid and amino acid sequence alignments. We use as a starting point the algorithm presented in[Mattei et al.(2014)] and improved it.In particular, modified versions of the Needleman-Wunsch[Needleman and Wunsch(1970)] and the Smith-Waterman[Smith and Waterman(1981)] algorithms, both using affine gaps, were implemented, respectively, for global and local sequence alignments.Moreover, a ClustalW-like[Larkin et al.(2007)] algorithm was implemented for multiple sequence alignment.Gap extension ad opening penalties were optimised using the same datasets described in[Mattei et al.(2014)]. These algorithms can compare and align RNA secondary structures,but can also take into account primary sequence information.
Results: Using the BEAR encoding we created a suite of aligning tools that outperforms available algorithms in computational complexity and execution time, while maintaining comparable performances.As a consequence, these algorithms are suitable for high-throughput applications, for example exhaustive searches for homologous RNAs in a database or suboptimal secondary structure predictions.This latter application may make it possible to face the problems and faults of the available programs for the inference of RNA secondary

structure.Implemented in Java, the BEAR suite is platform independent, and easy to use. A web-based version is also in preparation.

J60: Bryan Kowal, Akram Mohammed and Tomas Helikar. Building, simulating, and analyzing large-scale logical models of complex biological systems in a collaborative fashion with the Cell Collective

Abstract: Logical (e.g., Boolean) models provide a relatively easy and kinetic parameter-independent method of network model construction. However, the scale and complexity of biological systems is too large for one person or group to be able to construct computational models of substantial sizes and detail in a reasonable amount of time. The Cell Collective is an on-line platform (www.thecellcollective.org) designed to allow the world-wide scientific community to create large-scale, complex logical models collectively, in a crowd-sourcing fashion. Its user interface enables scientists to build and simulate models without manually creating complex mathematical equations (logical functions) or computer code, enabling laboratory researchers to contribute to the construction of these models. In addition, this platform allows scientists to simulate and analyze the models in real-time on the web, including the ability to simulate loss/gain of function and test what-if scenarios in real time to allow laboratory scientists to directly use these models as part of their day-to-day research. The Cell Collective currently contains hundreds of user-created models, and tens of published seed models that have been peer-reviewed in various journals. These seed models, containing nearly 2,000 components and 5,000 interactions represent biological and biochemical networks in organisms ranging from bacteria and viruses to yeast, flies, plans, and humans. Models in the Cell Collective are fully annotated within its wiki-like system, enabling researchers to track and discuss the biological evidence and assumptions used to construct each model. Because the biological context of models (and their components) in the Cell Collective is retained, they are readily available for the scientific communities (both biological and computational) to fully utilize and further build on them. Finally, the Cell Collective models are accessible and share-able not only within the platform, but they are also available for download in a number of open formats, including the recent SBML extension for qualitative models.

J61: Atefeh Lafzi, Saber Hafezqorani, Yesim Aydin Son and Hilal Kazan. Post-transcriptional regulation mediated by the interplay between RNA-binding proteins and miRNAs

Abstract: Post-transcriptional regulation (PTR) is mediated by the interactions of trans-acting factors with cis-regulatory sites in mRNAs. RNA-binding proteins (RBPs) and microRNAs (miRNAs) form the major classes of trans-acting factors in PTR. Recent studies have shown that each factor binds to hundreds of targets, and each mRNA is occupied by several factors. Also, RBPs and miRNAs are shown to function in coordination with each other [1]. Majority of previous research on PTR has focused on the effect of individual factors. In this study, we leveraged the recent explosion of PTR-related data to map both RBP and miRNA sites on mRNAs. We mapped RBP sites by taking into account i) motifs identified with RNAcompete [2], ii) peaks from existing CLIP data [3], iii) motifs from RBPDB database [4], iv) PhastCons conservation scores. To map miRNA sites, we took into account i) PicTar and TargetScan predictions, ii) Ago2 CLIP-identified peaks [4] iii) PhastCons conservation scores. In the next step, we analysed the mapped sites concurrently to detect potential interactions. These interactions could be competitive when there is overlap between the sites or cooperative when sites of two factors are located on each side of a stem (e.g. Pum and miR-221 [1]). We used our model to analyse the expression or stability of mRNAs that are

grouped based on the set of their constituent binding sites. In particular, we observed that mRNAs containing HuR sites with distinct secondary structure profile (both in terms of accessibility and potential competitive interactions) have distinct expression profiles upon HuR knockdown [5]. We also used our model to map sites of RBPs and miRNAs on the set of 3’UTR segments for which stability has been measured recently [6].

References

[1]Kouwenhove MV et al. MicroRNA regulation by RNA-binding proteins and its implications for cancer., Nat Rev Cancer 9 (2011), 644-656.

[2]Ray D, Kazan H et al. A compendium of RNA-binding motifs for decoding gene regulation., Nature 499 (2013), 172-177.

[3]Anders G et al. doriNA: a database of RNA interactions in post-transcriptional regulation. NAR (2012) D180-D186.

[4]Cook KB, Kazan H et al. RBPDB: a database of RNA-binding specificities. Nucleic Acids Res (2011) D301-308.

[5]Mukharjee N et al. Integrative regulatory mapping indicates that the RNA- binding protein HuR couples pre-mRNA processing and mRNA stability. Mol Cell (2011) 43(3):327-339.

[6]Zhao W, Pollack JL, Blagev DP et al. Massively parallel functional annotation of 3’UTRs. Nat Biotechnology (2014) 32(4):387-391

J62: Johannes Köster and Sven Rahmann. Massively parallel read mapping on GPUs with PEANUT

Abstract: We present PEANUT (ParallEl AligNment UTility), a highly parallel GPU-based read mapper with several distinguishing features. Most importantly, we introduce a novel q-gram index (called the q-group index) with small memory footprint built on-the-fly over the reads. The index can be accessed and built in a massively parallel way using prefix sums and population counts. To the best of our knowledge, this is the first feasible GPU-only implementation of a q-gram index. PEANUT allows to output both the best and all hits of a read, and outperforms other state-of-the-art CPU and GPU based read mappers in both fields. For the latter, 10x speedups could be achieved even on ordinary gaming GPUs.

J63: Matúš Kalaš, Sveinung Gundersen, László Kaján, Jon Ison, Steve Pettifer, Christophe Blanchet, Rodrigo Lopez, Kristoffer Rapacki and Inge Jonassen. BioXSD — A data model for sequences, alignments, features and measurements

Abstract: BioXSD has been developed as a data model and an exchange format for basic bioinformatics types of data: sequences, alignments, features and measurements. The BioXSD model is rich enough to enable loss-less capture of diverse data that would otherwise require use of multiple different formats and often even introduction of new formats for untypical features, classifications, or measurements. In BioXSD, an innovatively broad range of such experimental data, annotations, and alignments can be recorded in an integrated chunk of data, together with provenance metadata, documentation, and semantic annotation with concepts from ontologies of user's choice.

The last major release of BioXSD is version 1.1, and it has been released in form of a machine-understandable XML Schema (XSD). For the future, we plan to release BioXSD also in form of JSON Schema, XML Schema 1.1, and possibly also RelaxNG or even OWL. The aim is to share one data model that can be serialised into XML, JSON, RDF, or binary (EXI) on demand, while maintaining consistent and smooth validation, conversions, and parsing into objects for programming. The semantics of BioXSD is defined via SAWSDL references to EDAM (http://edamontology.org) and a number of other main ontologies.

BioXSD has always been an open community effort dependent on contributors, fans, and their needs, and can flourish only that way. BioXSD Schema, documentation, and examples are available at http://bioxsd.org.

J64: Jorge Alvarez-Jarreta, Gregorio de Miguel Casado and Elvira Mayordomo. PhyloFlow: A Fully Customizable and Automatic Workflow for Phylogeny Estimation

Abstract: Current phylogeny estimation systems such as DACTAL (1) or SATé (2) use fixed configurations and tools that make them suitable only for those problems in which that combination can provide an accurate solution. Out of that scope, a hand-made aggregation of individual methods has to be composed in order to get the desired phylogeny, like the workflow system we presented in 2011 (4).

PhyloFlow is a new framework based on a workflow designed extendable for a wide range of tasks in phylogenetic analysis. This system is specially intended to build large phylogenies, where most of the methods do not provide a solution. The workflow can be easily scalable to different phylogenetic estimation problems, the methods and stages already included can be fully customizable and once the user has set up the system, it will run automatically.

The first version we have built up covers two different systems: DACTAL (1) and our previous workflow (4). The workflow is divided in four different stages: i) automatic fetch of biological sequences (currently only GenBank (5)); ii) data preprocessing or multialignment (e.g. PRD from DACTAL (1), or Mafft (6)); iii) intermediate phylogeny estimation (e.g. model selection with PhyML (7) and bootstrapping); and iv) supertree stage (e.g. SuperFine (8)).

The system has being deployed in a cluster using HTCCondor and DAGMan (9). This way large phylogeny inference problems can be addressed, as is the case of reconstruction of the human mitochondrial DNA phylogeny, whose dataset is formed by up to 20000 sequences of 16569 bp.

Acknowledgement

Funding: The Spanish Ministry of Science and Innovation (Project TIN2011-27479-C04-01) and the Spanish Ministry of Education which provided J. Álvarez-Jarreta with a grant [FPU AP2010-1058] to carry out this investigation.

References

[1] K. Liu et al., “Rapid and Accurate Large-Scale Coestimation of Sequence Alignments and Phylogenetic Trees.” Science, vol. 324, pp. 1561, 2009.

[2] S. Nelesen et al., “DACTAL: divide-and-conquer trees (almost) without alignments.” Bioinformatics, vol. 28(12), pp. i274-i282, 2012.

[4] J. Álvarez et al., “Workflows with model selection: A multilocus approach to phylogenetic analysis.” PACBB 2011, vol. 93 of Advances in Intelligent and Soft Computing, pp. 39-47, 2011.

[5] D. Benson et al., “GenBank.” Nucleic Acids Research, vol. 38, pp. 46-51, 2010.

[6] K. Katoh et al., “MAFFT: a novel method for rapid multiple sequence alignment based on fast Fourier transform.” Nucleic Acids Research, vol. 30(14), pp. 3059-3066, 2002.

[7] S. Guindon and O. Gascuel, “A simple, fast, and accurate algorithm to estimate large phylogenies by maximum likelihood.” Systematic Biology, vol. 52(5), pp. 696-704, 2003.

[8] M. Swenson et al., “SuperFine: fast and accurate supertree estimation.” Systems Biology, vol. 61(2), pp. 214-227, 2012.

[9] J. Basney et al., “High throughput computing with condor.” HPCU news, vol. 1(2), 1997.

J66: Gonzalo Garcia Accinelli, Bernardo Clavijo, Sarah Ayling and Mario Caccamo. A systematic approach to plant genome assembly.

Abstract: Assembling plant genomes is difficult. No current technology produces the required sequencing data, and no current algorithm solves the problem with the data at hand. While generating a good quality draft of a mammalian genome (i.e. a human genome) is within the reach of current technologies and approaches, the same does not hold true for plant genomes which can be many times larger, consist of over 80% “repeat sequence” and may exhibit high ploidy levels. Plant researchers attempt to tackle these challenges using combinations of technologies and elaborate processing pipelines, with varying results. To date, a consensus on how best to design and execute plant genome sequencing and assembly projects has not yet emerged.

Within the TransPLANT project, we aim to generate a set of guidelines to help researchers when planning a plant genome project. We are surveying available data and processing strategies, focusing on data generation, QC, pre-processing, assembly and validation. We have selected a set of species ranging from the small genome of A. thaliana to the mid-sized A. sharonensis (7Gbp, 80% repetitions). We have also gathered and generated data from different technologies, including a range of Illumina paired-end libraries, Nextera Long Mate Paired libraries and Pacbio reads. We will explore different processing strategies for these datasets to show how the genome characteristics, data type and processing affect the final results, as assessed by different metrics.

The final result of this effort is to offer an open web-based resource that condenses all of the findings from these datasets and analyses, helping plant researchers to choose how to tackle their genome projects both in terms of data generation and analysis. We hope the methodological approach will be of continuing use to the community, with new datasets and analyses being added to keep it relevant and up to date. This will transform the challenging task of sequencing and assembling a plant genome into a well-informed process, with clear methodological steps and targets.

J67: Christophe Blanchet and Jean-François Gibrat. Bioinformatics cloud services of the French Institute of Bioinformatics

Abstract: Life science researchers, thanks to the continuous improvement of experimental technologies, face a deluge of data whose exploitation requires large computing resources and appropriate software tools. They simultaneously use many of the bioinformatics tools from the arsenal of hundreds available from the international community. Usually they combine their data with public data that are too large to be moved easily. So the computational infrastructure need to be tightly connected to public biological databases.

The French Institute of Bioinformatics (IFB) is the national infrastructure infrastructure which purpose is to provide bioinformatics core resources to the national and international life science research community (IFB is the French node of ELIXIR, the European bioinformatics infrastructure). Among the many tasks required to fulfill this goal, IFB must provide an IT infrastructure devoted to the management and analysis of biological data, in particular data generated by high-throughput technologies. This infrastructure will rely on sizeable hardware resources (high throughput computation, large storage capacity) and will provide access to high-quality developments in terms of software tools and databases. IFB consists of a network of more than 20 bioinformatics platforms gathered into six regional centers that span the French territory and a national hub called IFB-core (CNRS UMS3601). In particular, IFB-core is in charge of setting up and running the IFB academic cloud infrastructure that will be hosted in one of the French national HPC centers (IDRIS).

One important aspect of deploying a cloud for the life science will be to provide virtual machines (appliances) that encapsulate the many complex bioinformatics pipelines and workflows needed to analyze distributed life science data. At the IFB, we developed several bioinformatics services available as cloud appliances. A cloud appliance is a predefined

virtual machine that can be run on a remote cloud infrastructure. As cloud appliances have size usually of gigabytes, this is more efficient to moved them where the terabytes of biological data to analyse are stored instead of moving the data. But this requires to have at least some computing resources close to the stored data. We have created bioinformatics appliances providing, for example, a user-devoted Galaxy portal, a virtual desktop environment for proteomics analysis or a bioinformatics cluster with a lot of standard tools (BLAST, ClustalW2, R, Samtools, Bowtie, TopHat, etc.). Scientists can run their own appliances through a user-adapted web interface. Our cloud infrastructure is configured in such a way as to enable VMs to automatically connect to a local repository containing public databases, e.g., UNIPROT, EMBL, etc.
IFB is currently running an academic cloud infrastructure with the appropriate biological data and bioinformatics tools to meet the needs of the life science community.

J68: Foteini Pappa, Varvara Karagkiozaki, Paraskeui Kavatzikidou and Stergios Logothetidis. Development of conductive fiber-based scaffolds for tissue engineering and cellular uptake

Abstract: Tissue engineering is an interdisciplinary research area that combines the principles of engineering and life sciences towards the development of biological substitutes that restore, maintain and improve tissue function. Scaffold materials play a crucial role in nerve tissue engineering because they mimic the extracellular matrix (ECM), thus forming an appropriate microenvironment for the cells, allowing them to interact in vitro, efficiently. The “soft” nature of conductive polymers offers better mechanical compatibility with tissue, flexibility in surface functionalities and unique geometries for direct tissue regeneration. In this study, we incorporated a blend of the electrically conductive polymer, Poly (3, 4-ethylenedioxythiophene) Polystyrene sulfonate, (PEDOT: PSS) and the biodegradable polymer, polyvinyl alcohol (PVA), in order to fabricate conductive scaffolds, via electrospinning. The resulting structures had a nanofibrous non-woven morphology mimicking the ECM with the aim to manipulate cell growth and adhesion. To this end, a model cell-line L929 was seeded onto the engineered scaffolds in order to evaluate the cytotoxicity as well as the cellular attachment and proliferation for four different time periods (i.e., 6hr, 12hr, 3d and 6 d). Particularly, MTT cell proliferation/cytotoxicity assay was used for the evaluation of the cell toxicity, while imaging techniques such as scanning electron microscopy, fluorescence microscopy and methylene blue staining were incorporated in order to further evaluate the cell’s behaviour. Furthermore, a proper analysis of the surface nanotopography of the samples was carried out by Atomic Force Microscopy. Results indicate that the conductive scaffolds were found cytocompatible with promising cell adhesion and proliferation properties, thus providing good potential for their further utilization in nerve tissue engineering applications.