ECCB 2014 Accepted Posters with Abstracts

H: Biological knowledge discovery from data, texts and bio-images

H01: Louis Cronje, Rian Pierneef and Oleg Reva. Ribosomal RNA operons in horizontally transferred genomic islands – facts or artifacts?

Abstract: Genomic islands (GIs) in bacteria are commonly associated with their rapid environmental adaptation, including their pathogenicity. This amplifies the importance of the ability to accurately detect GIs in bacteria. The Pre_GI database is a web based platform which forms part of the SeqWord project and contains 26 744 GIs predicted by SWGIS program in 2407 bacterial chromosomes and plasmids. SWGIS is a composition based GI identifier and quite often ribosomal RNA operons (rrn) characterized by an alternative oligonucleotide usage are recognized by this program as GIs. Consequently, 4473 predicted GIs from 1344 replicons were initially discarded and deemed as false positives due to the presence of 16S RNA's as rrn operons are believed to be resistant to horizontal transfer. However, recent publications showed that the horizontal rrn transfer cannot be excluded. We investigated the discarded regions for the presence of genetic markers of true GIs and evaluated the distribution of such insertion sites.

H02: Sylvain Demey, Evelyne Begaud, Nahla Chaïbi, Jean-Mary Gallais, Elodie Raynal-Melchy, Loïc Talignani, Emmanuelle Helloin, Serge Casaregola, Anne Favel, Martial Briand, Florence Valence-Bertel, Marie-Laure Dardé, Isabelle Villena, Rémy Guyoneaud, Michaël Pressigout and Chantal Bizet. BRC-LIMS : application for the management of French Biological Resource Centres of Microorganisms

Abstract: Introduction: The operation and development of Biological Resource Centres (BRCs) make necessary the use of software to manage all data relating to acquisition, characterization, production, storage as well as distribution of biological resources. Currently, none of in-house developed software used by BRCs does not provide all the expected functions and that, for a maximum of types of biological resources.
Project Objectives: The purpose of this project, funded by a grant IBiSA GIS, was to develop a generic software, but flexible enough to meet the specific needs of each BRC, ensuring data traceability and compliance with regulations. This software must be distributable under a consortium agreement and interoperable with databases of national and international networks. Ten French microbial BRCs were project partners. In collaboration with the Biology IT Centre of the Institut Pasteur.
Topics / materials and methods: The application development was carried out by a service company in collaboration with BRC partners and the Biology IT Centre. The solution developed was composed of a public Web application (FBRCMi-catalog) and a private BRC-LIMS application. A BRC partners data recovery was also part of the specification.
The development was realized with the Java EE technologies (for the application part) and PostgreSQL 9 (for the database part).
The dialogue between BRC-LIMS for each BRC and FBRCMi-catalog, is done via a RESTfull webservice.
The maintenance and the new developments are now supported by the Biology IT Centre.
Main results: At the end of the project (start of production in January 2013), the BRCs had a

custom-designed tool for the management of microbial strains (BRC-LIMS), and a Web application FBRCMi-catalog (accessible via www.fbrcmi.fr) acting as common catalog for all BRCs participating in the project. These tools allow both the internal management of biological resources, their external visibility, and export data to other sites.
Conclusions: In conclusion, BRC-LIMS aims to be a tool used daily by BRCs to support users in their tasks and processes, sometimes complex for the management of biological resources. These software applications, allowing data exchange, facilitate the establishment of a national network of BRC microorganisms and potentially others biological resource networks in the future. It was designed to be generic, portable, interoperable and user-friendly as well as awareness of biodiversity represented in the BRCs to the scientific and industrial communities.

H03: Eero Lihavainen, Mikhail Kislin, Leonard Khirug and Andre Ribeiro. Automatic quantification of mitochondrial fragmentation from two-photon microscope images of mouse brain tissue

Abstract: Mitochondria are organelles that perform various vital functions in cells. Their dysfunctions are observed in many diseases, in particular neurodegenerative ones. Their morphology is shaped via the processes of fusion and fission; as a result, mitochondria can exhibit elongated, networked and fragmented morphologies. Recent research has shown that the fusion-fission dynamics, and consequently the morphology of mitochondria, is intricately related to their functioning; in particular, excess fragmentation of the mitochondrial network is associated with
many diseases.
Using recently introduced techniques of intravital fluorescence microscope imaging, it is possible to observe mitochondria in tissues of live animals. Combining such methods with animal models of diseases is likely to result in a better understanding of mitochondrial dynamics in disease. However, to exploit the images generated by such studies, novel computational tools are required.
Here, we present a method for the automatic quantification of mitochondrial fragmentation from two-photon fluorescence microscope images of mouse brain tissue, where mitochondria have been fluorescently labeled. The method learns a statistical model to predict a "fragmentation score", based on training data scored by experts, from features computed from a local region of a tissue image. We validate the method, and demonstrate its applicability with quantification results from experiments where mice underwent cardiac arrest.

H04: Melissa Mary and Gansel Xavier. LOINC & SNOMED-CT: usability and challenges to code identification tests and results for automated in vitro diagnostics systems

Abstract: In many countries simultaneous and similar analysis drove to the implementation of Electronic Medical Record (local to an hospital) and Electronic Health Record (national level) to store patient medical data such as patient allergy, clinical report and in vitro diagnosis reports.
This in the case in France with the Dossier Médical Patient (DMP) and in the USA through the Hitech Act.
These rises new interoperability challenges linked to syntax and semantic used to exchange underlying data. On one hand it calls for a common message structuration such as described by HL7 and definition of use case recommendation edited by the IHE (Integrating the Healthcare Enterprise). On the other hand it calls for a standardization of vocabularies used to describe information.
In case of in vitro diagnosis (IVD) the report should describe precisely the sample used but

also tests performed and associated results. As a part of the health care chain, IVD instruments should convey tests and results information with standard biomedical terminologies.
The purpose of this study is to analyze the possibility for microbiology IVD devices to use semantic interoperability recommendations to encode both observation tests and results. To perform this analysis we focused on both the IHE and HITECH Act recommendation which specified two distinct terminologies in case of laboratory reports : LOINC[1] (Logical Observation Identifiers Names & Codes) to express observation tests and SNOMED-CT[2] ontology (Systematized NOmenclature of MEDicine Clinical Terms) to express the observation results. Study is based on information given by four bioMerieux's systems, Vidas to detect antibodies against microorganisms, VITEK2 both identification and antibio-susceptibility testing card, VITEK-MS (Mass-Sepctroscopy) and Etest. Data are directly extracted from technical file reagents or instruments knowledge to be matched manually or automatically on those two terminologies.
We investigate the capabilities of those biomedical terminologies through the following questions :
1. Can LOINC accurately described automated test? microbial identification and antibio-suceptibility testing ? how exhaustive is it ?
2. Can SNOMED-CT accurately describe identification and antibio-susceptibility testing results obtained with those systems ?
3. How complete is SNOMED-CT compared to the reference bacterial nomenclature published by the ICSP[3] ?
[1]- LOINC : Logical Observation Identifiers Names and Codes ; http://loinc.org/
[2]- SNOMED-CT : Systematized Nomenclature of Medicine -- Clinical Terms ; http://www.ihtsdo.org/
[3]– ICSP (International Committee on Systematics of Prokaryotes) ; http://www.icsp.org/

H05: Matej Stano, Gabor Beke and Lubos Klucar. viruSITE - database of viral genomes

Abstract: Viruses are the most abundant biological entities and the reservoir of most of the genetic diversity on Earth. Despite they are the simplest biological forms bordering on the edge between living and non-living, their genomes are more diverse than in all domains of life. They vary in terms of their length (1.3 kb - 2500 kb), composition (ssRNA, dsRNA, ssDNA, dsDNA) and topology (linear, circular, segmented). Genomes of viruses were the first genomes to be sequenced. Currently, thousands of viral genome sequences are known. We present viruSITE, a database of genomes from viruses, viroids and satellites. It is a free database service hosted at the Institute of Molecular Biology SAS and accessible at http://www.virusite.org. viruSITE is a relational database using MySQL management system. Conventional web technologies like PHP, XHTML, CSS and JavaScript were used in the development of the user's interface. The web pages make use of Ajax in order to enhance their dynamics and minimize data transfer. In-house built genome browser phiGENOME is based on Adobe Flash technology. BLAST, Pfam search, MUSCLE, Circoletto and Circos are employed for data analysis and visualization. viruSITE database comprises information on all viral genomes and genes published in RefSeq database. Annotation extracted from RefSeq entries is complemented with information from UniProtKB, Pfam, GO, ViralZone and PubMed. This data integration makes viruSITE a comprehensive information resource in the field of viral genomics. Users can access the database content via web portal. 'Search' form allows searching database with keywords for specific search fields (virus name, host name, genome length, etc.). Browse section allows browsing through virus collection by names of viral species or by names of higher taxonomic groups. Search/browse results are presented in the form of two tables. The first one contains entries for viruses, the second entries for genes.

Entries for viruses contain information on virus taxonomy, host organisms, genome features (length, GC content, percentage of coding sequences, etc.) and references to other databases and bibliographic resources. Entries for genes comprise gene/protein names, position in genome, molecular weight and pI of proteins and references to other databases and bibliographic resources. User may select entries for further data retrieval or analysis. In the 'Retrieve' section of the database web portal, user can download (display or save) sequences of genomes, genes and proteins. Selected sequences can be also analysed in the 'Analyse' form. BLAST search, Pfam search and MUSCLE for multiple sequence alignment are at user's disposal. The 'Show' section contains phiGENOME genome browser which provides lucid and interactive graphical display of viral genomes. viruSITE is regularly updated. Current release (2014.2) contains genome sequences from 4043 viruses, viroids and satellites and 188702 entries for genes.
Supported by VEGA 2/0188/13

H06: Esther Schmidt, Oliver Pelz, Svetlana Buhlmann, Maximilian Koch and Michael Boutros. GenomeRNAi: A Phenotype Database for Large-scale RNAi Screens

Abstract: High-throughput RNA interference (RNAi) screening experiments have been established as a popular method for the systematic perturbation of gene expression. A wide variety of assays can be performed, resulting in the observation of loss-of-function phenotypes in many areas of biology. Data from these large-scale screens constitute a rich source for functional gene annotation.
The GenomeRNAi project aims to collect RNAi phenotype and reagent data and to provide them for data mining and comparisons. The database is populated by manual curation from the literature, or by direct submission by data producers, followed by curatorial review. To ensure consistent representation and comparability of the data, we have developed structured annotation guidelines. As of version 13, the database contains 194 experiments in human, and 183 experiments in Drosophila. Overall, it holds more than 950,000 individual gene-phenotype associations. Furthermore, it contains information on more than 400,000 RNAi reagents, along with a quality assessment with respect to their efficiency and specificity. GenomeRNAi is publically accessible at www.genomernai.org. It allows searching by gene, reagent, or phenotype; browsing and downloading screening experiments; or visualizing phenotypes and reagents in their genomic context, powered by a DAS server. A “frequent hitter” page indicates genes that are often associated with a positive phenotype. Another feature allows the overlay of genes sharing the same phenotype onto known gene networks provided by the String database (www.string-db.org), facilitating further exploration in a functional context. GenomeRNAi data are well inter-linked with external resources, providing e.g. mutual links with FlyBase, GeneCards and UniProt. GenomeRNAi functional data have also been integrated into the FlyMine tool.
The GenomeRNAi website features a private space for data producers for managing their submissions. They can download the submission template, upload and display their data, share them with colleagues or reviewers via a private URL, and finally submit them for inclusion in the next public release of the database.
A recent addition to the website is the screen comparison tool, implemented as a heat map-like overview over the genes showing phenotypes in different screens. This tool has been made available as a beta version to solicit user feedback during its development.

H07: Michael Lenz, Daniela Malan, Joana Frobel, Wolfgang Wagner, Philipp Sasse and Andreas Schuppert. A two-scale gene expression landscape for the characterization of in vitro differentiated cells

Abstract: Gene expression microarray analysis usually concentrates on the interpretation of differences between phenotypes, e.g. healthy and diseased, or on the development of biomarkers for the detection of specific cell types or disease states. Such rather local analyses can be complemented by global approaches, locating new data, e.g. from in vitro differentiated cells, in the high dimensional gene expression space, evaluating their similarity to various types of well characterized tissues or cell types. Here we propose a two scale gene expression landscape together with robust mapping algorithms to tackle this task.
Based on a previously published dataset* of 5372 samples from 369 different tissues or cell lines, we created a two-scale gene expression landscape extending the low dimensional principal components space* with a residual space consisting of 369 tissue specific expression patterns.
In vitro differentiated cells are then mapped to these two spaces in order to determine their similarity to the various tissues or cell lines.
We show that through the use of appropriate reference datasets and robust mapping algorithms, the two-scale expression landscape can be used to analyze data across microarray platforms. We apply the developed method to various in vitro differentiated cells and discuss the observed similarities and differences to the corresponding in vivo cell types. The mapping approach confirms the proper differentiation of mesenchymal stromal cells, derived from induced pluripotent stem cells, and is able to detect differences between cardiomyocytes that were differentiated and purified by different protocols. The method can thus be used as a quality control and optimization criterion for differentiation protocols.
* Lukk et al. A global map of human gene expression. Nat Biotech, 28:322-324 (2010).

H08: Anne Mai Wassermann. Integrating historical biological data for small molecule-target predictions

Abstract: With the renaissance of phenotypic screens in drug discovery, computational approaches are increasingly needed to elucidate targets of small molecules that show a desired phenotypic effect but have an unknown mechanism-of-action (MoA).
Fortunately, modern screening technologies have led to a growing body of small molecule bioactivity data in both the public and corporate domain which can be leveraged to generate MoA hypotheses.
We introduce a method that encodes small molecules based on the activity patterns that they have shown in more than 200 biochemical and cellular assays run at Novartis over the past decade. In target elucidation efforts, we match activity patterns of more than 2,500 natural products and drugs to an MoA-annotated compound reference panel and use a guilt-by-association principle to predict their targets. Our large-scale target prediction reveals substantial differences in the protein targets modulated by the two compound classes and gives guidelines how to target the druggable genome.
In vitro experiments confirm 48 novel protein interactions for marketed drugs; in many instances, newly discovered targets are able to explain a drug's efficacy or side effects. Importantly, our approach reveals mechanisms-of-action that cannot be inferred from the chemical structure of a molecule.
This work emphasizes the importance of leveraging existing biological data in the investigation of small molecules. The massive amount of screening data released both in corporate and public domains gives us the outstanding opportunity to generate target hypotheses based on what we already know about the biological actions of a compound. Ultimately, this should lead to a more complete picture of the perturbations that a compound induces on a complex biological system and a better assessment of its efficacy and safety in drug development.

H09: Catherine Kirsanova, Ugis Sarkans and Gabriella Rustici. Cellular Phenotype Database with Ontology assisted data browsing

Abstract: The Cellular Phenotype database stores data derived from high-throughput phenotypic studies, it is being developed as a part of the Systems Microscopy Network of Excellence project.
Within this project, we have developed a public data repository, which provides access to systems microscopy data for the broader research community, facilitates development of analytical methods for this field and offers standards for enabling aggregation with other molecular biology data infrastructure.
The aim of the Cellular Phenotype database is to provide easy access to phenotypic data and facilitate the integration of independent phenotypic studies with the help of ontology assisted data browsing.
Methods and Algorithms: The Cellular Phenotype database is built using a flexible, open source document-oriented NoSQL database management system MongoDB and allows for aggregation of data in different formats.
Leveraging on MongoDB's flexibility and power, we are building an infrastructure for data integration that enables development of new algorithms for data analysis as well as improving our understanding of how cellular systems function.
Results: The Cellular Phenotype database accepts data from high-throughput phenotypic studies. Such studies allow screening living cells under a wide range of experimental conditions and give access to a whole panel of cellular responses to a specific treatment. Substances like small molecules and peptides, or techniques like RNA interference (RNAi), can be applied to look at the effects, or phenotypes, that such substances induce in cells, with the aim of elucidating novel gene function, as well as screening compounds for desirable therapeutic effects.
Through Cellular Phenotype database interface, users can search for a gene of interest, either by its symbol / Ensembl ID or by its ontology attributes, and retrieve the phenotypes observed across independent phenotypic studies.
Also, users can search for a phenotype or phenotype set, either by its study specific name(s) or by its corresponding ontology term(s), and retrieve the reagents that have caused such phenotype(s) as well as the associated target genes. Information about specific reagents can be obtained when querying by a reagent ID.
Alternatively, users can explore all datasets loaded in the database by browsing the phenotypes — also either by their study specific names or by their corresponding ontology terms. Users can also search for studies by keywords.
Data in the repository are aggregated in a way that users can compare the results (say, observed phenotypes) within a given study or across different studies.

H10: Lisa M Breckels, Sean Holden, Kathryn S Lilley and Laurent Gatto. A Transfer Learning Framework for Organelle Proteomics Data

Abstract: Organelle proteomics, or spatial proteomics, is the systematic study of proteins and their assignment to subcellular niches including organelles. It is a field of rapidly growing importance as many diseases result from protein mislocalisation. The knowledge of subcellular localisation of proteins is extremely desirable to biologists, as it can assist elucidation of a protein's role within the cell, as proteins are spatially organised according to their function and specificity of their molecular interactions. There exist a number of sources of information available with which to accurately assign a protein to its subcellular compartment. These data sources encompass data produced from experimental high-throughput mass spectrometry (MS)-based methods (e.g. [1, 2]) and freely available in silico

data such as amino-acid sequence representations and Gene Ontology information. MS approaches to protein-organelle association are a recent development and to date these methods have largely employed straightforward supervised learning to map unannotated abundance profiles to known protein-organelle associations [3]. Here, we use machine learning in a data fusion and inductive transfer framework [4] to simultaneously exploit several sources of information available with which to improve upon the classification of experimental subcellular localisation predictions. We show that simple data fusion of these different, but complementary, sources of information does not yield an improvement over using experimental data alone. However, we find that weighted ensembles, both data source specific and organelle class specific are a promising avenue for not only subcellular localisation prediction but for data exploration and quality control. Finally, we show when given data from a high quality MS experiment, integrating data from a second more plentiful auxiliary data source directly in to classifier training and classifier creation results in the assignment of proteins to organelles with high generalisation accuracy. Furthermore, we show that our data fusion methods outperform a single classifier trained on each single data source alone. A new transfer learning framework for predicting protein localisation prediction that uses a novel machine learning methodology to unify multiple sources of information is proposed. This methodology forms part the Bioconductor pRoloc [5] suite of computational methods available for organelle proteomics data analysis.

References:
[1] Dunkley, T. et al. PNAS, 2006, 103: 6518– 2. [2] Tan, DJ. et al., J Proteome Res. 2009, 8, 2667-78. [3] Trotter, M.W.B., et al. Proteomics 2010. [4] Wu, P and Dietterich, TG, Proc. 21st Int'l Conf. Machine Learning. 2004, 110. [5] Gatto, L et al. Bioinformatics 2014, 30(9):1322-4.

H11: Tsubasa Ogawa, Kenji Etchuya and Yuri Mukai. Studies on factors and prediction of protein palmitoylation using a back propagation artificial neural network (BP-ANN)

Abstract: Bio-synthesized proteins are modified by a variety of post-translational modifications (PTMs). PTMs are related to functional expressions and controls of proteins. Lipid modification is one of the PTMs which can increase protein hydrophobicity and affinity for lipid membranes. Myristoylation, palmitoylation, prenylation and GPI-anchor are known as general lipid modifications. Palmitoylation is a lipid reversible modification in which palmitic acid is modified on a cysteine residue. Palmitoylation plays important roles in regulating the protein-lipid membrane and protein-protein interactions, switching protein functions, intracellular localization, structural stability as well as other functions. As mentioned above, the elucidation of the palmitoylation mechanism is thought to be important in understanding protein function and interaction.

Many palmitoylation prediction tools have been developed recently. While previous tools can predict palmitoylation using machine learning based on amino acid propensities, they have difficulty understanding physicochemical factors around the palmitoylation site recognized by palmitoyl-acyl-transferase. Therefore in this study, a palmitoylation prediction method was developed using back propagation artificial neural network (BP-ANN) combined with amino acid physicochemical characteristics. Protein palmitoylation consensus sequences and physicochemical factors were examined based on the results of this prediction.

The dataset of palmitoylation sites was extracted from the sequence data of proteins obtained in UniProtKB/Swiss-Prot Release 2014_01. By calculating the amino acid propensity around palmitoylation sites in the dataset, the position specific score (PSS) was estimated. According to the PSS, hydrophobic amino acids were appeared frequently around the palmitoylation site. A palmitoylaton prediction method was developed using BP-ANN, and some physicochemical characteristics of amino acids were combined with this method. The

prediction accuracy depended on the combination of amino acids characteristics, thus the important residues for palmitoylation could be detected by using the prediction accuracy differences.

H12: Afaf Benhouda, Mouloud Yahia and Hachani Khadraoui. Gastroprotective activity of Umbilicus rupestris leaf extract in experimental animal

Abstract: The main objective of this study is to evaluate the antiulcer activity of extract of Umbilicus rupestris leaf extract against ethanol-induced gastric mucosal injury in rats [1]. Four groups of Wistar rats were pretreated, respectively, with distilled water; omeprazole 20 mg/kg; and 100, and 200 mg/kg U.rupestris leaf extract 30 min before oral administration of absolute ethanol to generate gastric mucosal injury. After one hour later, the rats were sacrificed and the ulcer areas of the gastric walls were determined. Gross evaluation has revealed that the negative control rats exhibited severe mucosal injury, whereas, pre-treatment with U.rupestris leaf extract resulted in significantly less gastric mucosal injury and flattening of the mucosal folds[2]. Histological studies of the gastric wall that the pre-treated with U.rupestris leaf extract where there was marked gastric protection along with reduction or inhibition of edema and leucocytes infiltration of the submucosa.

H13: Jaroslav Budis, Rastislav Hekel, Gabriel Minarik and Tomas Szemes. Application and web service for functional annotation of variants

Abstract: The rapid advance of the high-throughput sequencing technologies has significantly accelerated discovery of new genetic variants related to heterogenic rare disorders as well as common diseases and phenotypic traits. However, common large-scale screening processes lead to identification of a large number of candidate variants, of which only a small part is related to the studied phenomenon. The essential step of the analysis is therefore filtering and prioritizing based on their functional impact.
Proper evaluation of the functional effect requires a rich annotation of the candidate variants with attributes representing their conservation among population, affected gene and an impact on functionality of a translated protein. Annotation data are scattered across various databases, which makes manual annotation a time-consuming and tedious process, even for a small subset of selected attributes.
To facilitate the annotation process, we developed application with graphic interface called Variant Annotation Analyzer (VAA). Application acts as a thin client which utilizes retrieving of variant attributes from various web sources. Application also supports follow-up prioritizing process with filtering and sorting operations and export to the common tabular formats. Modular architecture based on plugins allows easy extension with a newly emerged functionalities and additional attribute providers.
The main attribute provider for the VAA application is the web service called Variant Annotation Service (VAS). The service acts as a front-end for the database of non-synonymous SNP's functional predictions, dbNSFP and provides more than 100 attributes collected from various variant and gene databases, conservation scores and results of functional prediction tools. The use of the web service is not limited to the VAA application and is fully open to any academic use.
The VAA application and the VAS web service provide researchers with rich and automated annotation of variants in a fraction of time compared to manual annotation. With the option of further prioritization it provides a powerful tool for fast identification of potential candidate mutations among loads of irrelevant variants.

H14: Nikolay Samusik, Brice Gaudilliere, Gabriela Fragiadakis, Robert Bruggner, Martin Angst and Garry Nolan. Fast and accurate mapping of phenotypic space in single-cell data with X-shift

Abstract: Highly parameterized mass cytometry allows routine measures of up to 45 protein markers from single cells and can provide a system-level overview of various immune processes (via staining of cells with comprehensive panels of mass-tagged antibodies against surface markers and intracellular network response states). Information about hundreds of thousands of cells can be collected from a single instrument run, resulting in information-rich datasets. However, comprehensive manual analysis of cell populations in multidimensional data by means of sequential gating on biaxial plots can be an intractable task given the immense number of possible gating strategies that would be required to examine every major and minor cell subset in a 45 dimensional space. Thus there is a need for an algorithm that is able to reliably and deterministically map all major cell types. Automated density-based clustering reasonably mirrors human intuition driving manual hand-gating, but can also accommodate cells with multidimensional features, without requiring any parametric assumptions about the population sizes and distribution shapes. However, efficient estimation of density in large and multidimensional datasets can be challenging.
To address this, we created a new density-gradient clustering method (X-shift) that is based on k-nearest neighbour density estimate.. Our simulations show that KNN-DE is able to accurately estimate the density of normal as well as long-tailed distributions in multi-dimensional space (upward of 100 dimensions). We developed a new fast and exact KNN search strategy with sub-quadratic complexity that enables computation of density estimate on extra-large datasets (several millions of data points) in reasonable time. We used X-shift to analyse a human CyTOF dataset containing 2.7M PBMCs from 26 human patients that were measured on 21 surface markers. X-shift clustering produced a small and stable number of clusters and accurately reproduced the expert hand-gating on most well-known cell populations, with the exception of a few cases where the population-defining marker distribution was not clearly bimodal. X-shift also found some of the extremely rare but genuine cell types in this dataset that were missed by the gating expert, including double-negative T-cells (1.7%) or CD7+ pDCs (0.04%). X-shift clusters and their relationships can be conveniently visualized using either minimum spanning trees or automatically reconstructed phylogenies, where each population is treated like a separate species and surface markers serve as features. X-shift clustering is available a part of a GUI tool VorteX(http://sites.google.com/site/vortexclustering) that enables comprehensive cluster analysis and visualization.

H15: Peter Ebert, Christoph Bock and Thomas Lengauer. A computational approach towards cross-species epigenomics

Abstract: Compared to the genomic DNA sequence, epigenomes are highly dynamic and consist of various biochemical marks. Characterizing the complete epigenome of any single organism is therefore a formidable task. As we have learned from genomics already, comparative analyses involving model organisms are one of the main sources for new insights into cellular processes. The hypothesis underlying our work is that this is true also for epigenomics: A computational genome-scale estimate of the epigenetic state in several species can help scientists to focus on a small set of interesting regions before doing wet lab work, thus reducing the amount of expensive and laborious experimental assays. Our computational approach towards comparative epigenomics employs a dual strategy: (i) we map epigenetic data, such as signal tracks from histone ChIP-seq experiments, between corresponding cell types of different species following the hypothesis that sequence

conservation fosters a reasonably stable epigenetic signal in the respective genomic region. To complement the mapping, we use machine learning techniques to train statistical models based on sequence-derived features to estimate the presence of an epigenetic mark, e.g. a certain histone modification, at a specified location. The combined information of mapping and prediction is then used to characterize genomic regions in the species of interest solely based on epigenome data of a reference species and genome data of the target species. Our preliminary results indicate that the mapping alone can only provide a rough outline of the true epigenetic signal. Nevertheless, in an internal validation study, where we predicted gene expression levels using inferred epigenetic data, we observed performance increases compared to the naïve baseline model, which was trained on genomic features only. We are now extending our approach to include data from different sources like ENCODE or BLUEPRINT to improve robustness and to assess model limitations in the presence of cell-to-cell variation and cross-species differences.

H16: Kota Hamada, Kenji Etchuya and Yuri Mukai. Influence of signal-peptide sequences on subcellular location of mature proteins

Abstract: Many proteins are biosynthesized in ribosomes and then transported to translocons by endoplasmic reticulum (ER)-targeting signals. The ER-targeting signals in the N-terminus, called signal-peptides, are separated from mature proteins by signal-peptidase after being transported to the ER and degradated in the lipid bilayer. Many varieties of signal-peptide sequences exist and have regions with high hydrophobicity, however, the reasons why signal-peptide sequences are not unique have not been clarified.
The ER is the starting point for transporting mature proteins to each cell organelle after being separated from the signal-peptide. Mature proteins which were transported into the ER by translocons, then located to the appropriate subcellular localization based on the transport mechanism of each protein.
Previous results about amino acid propensities around hydrophobic regions in signal-peptides implied that the subcellular localization of mature proteins was thought to be influenced by the Pro residues upstream from hydrophobic regions in signal-peptides.
Therefore, the influences of amino acid propensities on the subcellular localization of mature proteins around hydrophobic regions and cleavage sites were investigated in this study. The sequence and subcellular localization information of mature proteins of mammalian proteins which have signal-peptides were first extracted from UniProt Knowledgebase/Swiss-Prot release 2013_04 and release 2014_01. The amino acid propensities around the hydrophobic regions were especially examined in the extracted sequences. The position-specific scores were calculated by each signal-peptide sequence and evaluated by the cross-validation test with various calculation domains. The amino acid propensities, the physicochemical properties and the hydrophobicity in extracted sequences, classified into groups according to the subcellular localization of mature proteins and cleavage residues. In addition, the recognition factors of signal-peptidase were investigated in this study. The influence to the subcellular localization on physicochemical properties of signal-peptide sequences was also analyzed and reported in this study.

H17: Kenji Etchuya and Yuri Mukai. Structural Characteristics of Fuc Modification Sites in Mammalian Proteins

Abstract: Most proteins in eukaryotic cells are modified by post-translational modifications (PTMs) in the endoplasmic reticulum and the Golgi body and grow into mature proteins after going through PTM processes. Glycosylation, one form of PTM, is known to affect protein folding, protein functions and enzyme activities. Through O-glycosylations, motif residues are

attached to variations sugar chains (usually Ser or Thr) with glycosyltransferases in the Golgi body. After undergoing glycosylation, they become membrane binding proteins and secretory proteins. Each sugar type plays different roles in living cells.
While proteins have a motif residue, sugar chains are not always modified at the modification site, therefore consensus sequences except motif residues have not been clarified completely. Some characteristics based on each sugar type have been extracted by bioinformatic techniques from experimental evidence, and prediction tools using weak characteristics have been developed enabling the detection of new O-glycosylation sites. The prediction tools already published could predict only GalNAc, Glc and GlcNAc modification sites. Combining protein structure information with the current method for predicting protein glycosylation will enable more accurate discrimination of sugar types. Therefore, the development of new prediction tools which can predict new sugar types of O-glycosylation with high accuracy is expected in the near future.
In this study, to find structural characteristics based on each sugar type, the three-dimensional coordinate data of atoms in amino acids around the O-glycosylation site was analyzed. Mammalian protein data which has crystal structure information was extracted from the Uniprot KB/Swiss-Prot 2013_04. Entries that have O-glycosylation sites on “SER” or “THR” in the crystal structures were extracted from the Protein Data Bank (PDB) release 2012_04, and were classified into different sugar types. Secondary structure information of each entry with eleven amino acids around the O-glycosylation sites was extracted from the PDBFINDER2 which had secondary structure information defined by the DSSP algorithm. Through propensity analysis of the protein secondary structures in Fuc modification sites, the correlation between sugar type and protein secondary structure in O-glycosylation was reported.

H18: Emad Elsebakhi, Rashid Al-Ali, Mohamed-Ramzi Temanni, Abdou Kadri, Radja Badji and Rawan Alsaad. DECISION SUPPORT AND OUTCOME PREDICTION WITHIN SIDRAiTrip TRANSLATIONAL RESEARCH INFORMATICS PLATFORM

Abstract: Personalized medicine is aiming at treating patients with respect to their individual genome. In the past decade, the advance in sequencing technology coupled with a reduction in cost per genome led to a large volume of heterogeneous genomic data generated per patient. Bioinformatics and clinical informatics play a key role in deciphering the genomic data and identifying structural variants associated with disease. To address some of these diverse needs, a data analytics framework that facilitates data mining and machine learning is proposed and integration strategies of both phenotypic and genotypic data that leads to biological knowledge discovery is incorporated.
Since biomedical research is becoming more data-intensive; as physicians and researchers are generating increasingly complex and diverse biomedical data; an important challenge in biomedical informatics field is to couple several levels of genomic data to build decision support systems capable of helping physicians in decision making and outcome prediction in order to bring research finding to the bedside and provide a tailored treatment for patients. Several machine learning approaches were used with clinical data, however not all of these methods apply to genomic data nor can cope with integrating several large-scale heterogeneous biomedical sources.
Sidra Medical and Research Center is a new all-digital academic medical and research center aiming at providing a high standard of care for women and children as well as a high quality biomedical research in Qatar. Sidra Integrated Translational Research Informatics Platform (SIDRAiTrip) is an integrated platform for storing, analyzing and interpreting biomedical data and provides researchers and scientists with a set of features to better interact with the data and answer research questions in an intuitive and easy manner.

The implementation of SIDRAiTrip focuses on integrating all possible sources of data within the hospital setting (clinical data, genomic data, electronic medical records, and registries) in a centralized data warehouse. Furthermore, the integration layer allows the incorporation of the analytics framework and the various tools. Empowering SIDRAiTrip, tools for excellent query, data analysis and visualization capabilities are been implemented. In addition, export functionalities are developed allowing for advanced analysis, and knowledge discovery data mining predictive mdeling.
In this work, the analytics framework within SIDRAiTrip is presented; the framework implements state of the art big data analytics, machine learning, and computational intelligence predictive modeling tools as well as tools used in Omics data analysis. The implementation of key features and the learning processes are briefly proposed. An integrative design to combine different data sources as well as an extensive benchmark of these tools on multi-Omics data is presented.

H19: Nikolaos Papanikolaou, Georgios Pavlopoulos, Evangelos Pafilis, Theodosios Theodosiou, Reinhard Schneider, Venkata Pardhasaradhi Satagopam, Christos Ouzounis, Aristides Eliopoulos, Vasilis Promponas and Ioannis Iliopoulos. BioTextQuest+: A knowledge integration platform for literature mining and concept discovery

Abstract: The iterative process of finding relevant information in biomedical literature and performing bioinformatics analyses might result in an endless loop for an inexperienced user, considering the exponential growth of scientific corpora and the plethora of tools designed to mine PubMed and related biological databases. Herein, we describe an automated approach to bridge processes such as bioentity recognition, functional annotation, document clustering and data integration towards literature mining and concept discovery via the web-based BioTextQuest+ platform. BioTextQuest+ enables PubMed and OMIM querying, retrieval of abstracts related to a targeted request and optimal detection of genes, proteins, molecular functions, pathways and biological processes within the retrieved documents. The front-end interface facilitates the browsing of document clustering per subject, the analysis of term co-occurrence, the generation of tag clouds containing highly represented terms per cluster and at-a-glance popup windows with information about relevant genes and proteins. Moreover, to support experimental research, BioTextQuest+ addresses integration of its primary functionality with biological repositories and software tools able to deliver further bioinformatics services. The Google-like interface extends beyond simple use by offering a range of advanced parameterization for expert users.
Availability: The service is accessible at: http://bioinformatics.med.uoc.gr/biotextquest

H20: Yan Zhang, Isabel Riba-Garcia, Richard Unwin, Henning Hermjakob and Andrew Dowsey. Streaming Visualisation for Raw Mass Spectrometry Data and Results Based on a Novel Compression Algorithm

Abstract: Mass Spectrometry (MS) has become a pervasive technique for the analysis of biological compounds. Vast amount of data are generated using high-throughput LC-MS (Liquid Chromatography–MS). To keep pace with this data explosion, great endeavours have been made in the bioinformatics field using more intricate and automated analyses. There is a danger that the process becomes more and more opaque and inaccessible to MS practitioners. It is therefore vitally important that efficient visualisation tools are available to facilitate quality control, verification, validation and interpretation of those big data: raw MS data and MS analyses.
MS data converted to the standards-compliant Proteomics Standards Initiative mzML format are organised as a contiguous list of raw spectra. It is therefore fast to recall individual spectra

due to the indexing scheme and their relatively small size. However, generating a 2D overview ('virtual gel') of an LC-MS dataset requires every single datapoint to be loaded, which takes significant time and memory space. Also, 2D panning and zooming for MS data stored in those formats is inefficient. To tackle these issues, we leverage a novel compression algorithm developed within our seaMass framework. This transform compressor employs a complete set of separable 2D multiscale cubic B-spline basis functions as a sparse domain; this models the raw LC-MS data as the non-negative weighted sum of a collection of overlapping tensor-product B-spline basis functions. The decompression operation is simply the sum of those B-spline building blocks and can be conducted on the fly. This scheme is able to achieve lossless compression with high compression rates. By adjusting the shrinkage parameter, MS data can be de-noised and compressed with compression ratio higher than 200:1.

In MS data, peaks are cognitively most important. To organise compression coefficients both spatially and ordered by intensity, an Octree data model is employed. Octrees are most often used to represent objects in 3D spaces. Retention time, mass-to-charge ratio, and quantised coefficients are selected as the three dimensions for our MS Octree. This Octree creates a hierarchy of Axis Aligned Bounding Boxes, for visualisation which is used to cull away coefficients outside of view, and retrieve coefficients in a sorted order by visual importance. Data saved in the Octree can be streamed in the order of their importance, and the visualisation iteratively refined: early approximate reconstructions will permit fast response to initiation, panning and zooming, while later reconstructions will reveal further details. Underpinned by the compression scheme and the efficient data organisation model, a prototype system has been developed to stream MS datasets at any desired resolution. This visualisation engine is fused with detailed annotations of identification and quantification results, therefore providing a complete dissemination system for the mass spectrometrist.

H21: Gaston Mazandu and Nicola Mulder. Information Content-based Gene Ontology Functional Similarity Measures: How good are these measures?

Abstract: The current increase in Gene Ontology (GO) annotations of proteins in the existing genome databases and their use in different analyses have fostered the improvement of several biomedical and biological applications. To integrate this functional data into different analyses, several protein functional similarity measures based on GO term information content (IC) have been proposed and evaluated, especially in the context of annotation-based measures. In the case of topology-based measures, each approach was set with a specific functional similarity measure depending on its conception and applications for which it has been designed. However, it is not clear whether a specific functional similarity measure associated with a given approach is the most appropriate, given a biological data set or an application, i.e., achieving the best performance compared to other functional similarity measures for the biological application under consideration.

We show that, in general, a specific functional similarity measure often used with a given term IC or term semantic similarity approach is not always the most 'adequate' for different biological data and applications. We have conducted a performance evaluation of these different functional similarity measures using different types of biological data in order to infer the best functional similarity measure for each different term IC and semantic similarity approach. The comparisons of different protein functional similarity measures should help researchers choose the most appropriate measure for the biological application under consideration.

H22: Arthur Tenenhaus, Vincent Guillemot, Vincent Perlbarg, Andigoni Malousi, Justine Guégan and Ivan Moszer. RGCCA: a versatile tool for the analysis of multiblock and multigroup datasets

Abstract: A challenging problem in multivariate statistics is to study relationships between several sets of variables that are measured on the same set of individuals (called blocks throughout this abstract). In statistics, this paradigm is referred to as "multiblock data analysis", whose main objective is to identify variables in each block that contribute to block relationships. For instance, neuroimaging is increasingly recognized as an intermediate phenotype to understand the complex path between genetics and behavioural or clinical phenotypes. In this imaging-genetics context, the goal is primarily to identify a set of genetic biomarkers that explains the neuroimaging variability which is linked to a modification of the behaviour. It is therefore crucial to perform multiple experiments (e.g. SNPs, functional MRI, behavioural data) on a single set of patients, resulting in a complex "multiblock data set". The joint analysis of such datasets becomes possible with the use of Regularized Generalized Canonical Correlation Analysis (RGCCA) [Tenenhaus and Tenenhaus, 2011].
In contrast, multigroup data analysis concerns the analysis of one set of variables observed on a set of individuals who are distributed amongst groups. In this setting, the main objective is to identify shared patterns of correlation across groups. For instance, it is essential to reliably detect the metabolic or functional changes associated to the onset of disease symptoms. Namely, it is argued that a family of neurodegenerative diseases share common molecular mechanisms that are suspected to "activate their onset and progression. In this "cross-disease context", the same set of variables (e.g. voxels in a neuroimaging modality) are collected in different studies and one looks for common patterns of correlation across diseases. Thus, developing methods able to fit the multigroup structure/paradigm of the data becomes crucial. It appeared that RGCCA, which was initially designed to be a unifying approach for multiblock data analysis can be extended to the multigroup context [Tenenhaus and Tenenhaus, 2014], through a simple adaptation of the RGCCA optimisation problem.
The versatility and usefulness of our approach is illustrated on real multiblock and multigroup datasets, generated notably in the frame of multimodal studies run on neurodegenerative diseases at the IHU-A-ICM, whose integrative analysis is handled by the bioinformatics/biostatistics core facility.
A. Tenenhaus and M. Tenenhaus, "Regularized Generalized Canonical Correlation Analysis," Psychometrika, (76), 257–284, 2011.
A. Tenenhaus and M. Tenenhaus, "Regularized generalized canonical correlation analysis for multiblock or multigroup data analysis," European Journal of Operational Research, (238), 391–403, 2014.

H23: Abhishek Dixit and Richard Dobson. CohortExplorer: A generic application programming interface (API) for entity attribute value database schemas

Abstract: Most electronic data capture tools developed to collect and store clinical data from participants recruited into studies are based on generic Entity-Attribute-Value (EAV) database schemas which enable rapid and flexible deployment in a range of study designs. The drawback to such schemas is that they are cumbersome to query with structured query language (SQL). The problem increases when researchers involved in multiple studies use multiple electronic data capture tools each with variation on the EAV schema. We present a tool, CohortExplorer, which will 'plug-n-play' with EAV schemas, enabling the easy construction of complex queries through an abstracted interface. CohortExplorer allows easy and rapid exploration of data and metadata stored under EAV schemas. CohortExplorer is written in Perl programming language and uses the concept of SQL abstract which allows the

SQL query to be treated like a hash (key-value pairs). To demonstrate the utility of the CohortExplorer tool, we show how it can be used with the popular EAV based frameworks; Opal (OBiBa) and REDCap. The tool is available under a GPL-3+ license at the following URL: https://metacpan.org/pod/CohortExplorer.

H24: Alexandre Angers-Loustau, Mauro Petrillo, Alex Patak and Joachim Kreysa. GMO-Scan: a tool for fast identification of transgenic elements in DNA sequences.

Abstract: The Joint Research Centre (JRC) of the European Commission is host to the European Union Reference Laboratory for Genetically Modified Food and Feed (EU-RL GMFF). The EU-RL GMFF maintains the Central Core DNA Sequences Information System (CCSIS), an in-house database containing Genetically Modified Organisms (GMO) sequences, either submitted to the EURL-GMFF by industry or identified within public databases (EMBL, Genbank, Patents, etc).
Sequences of the CCSIS database are extensively annotated and the whole set of metadata is stored into a relational database. Briefly, each GMO event is composed of several different genetic elements, i.e. specific sub-regions of the transgenic cassette, like promoters, terminators or coding sequences. Elements are in turn classified and grouped, i.e. elements sharing the same origin and/or specific sub-regions are assigned to a specific element tag. An automatic script-based procedure has been developed that, for each of the element tags, retrieves the sequences of the element tag members, aligns them by running ClustalW and generates a consensus from the alignment. All the consensus sequences are then transferred as a multi-FASTA file to a local file server repository, available for automatic retrieval.
This information is then used by the GMO-Scan application, a web-based tool developed at the EU-RL GMFF, which allows a similarity comparison of a given unannotated DNA sequence with the library of GMO-elements' consensuses using Megablast, a tool developed at NCBI and installed locally. The platform is built on three main servers: a server that generates the GMO elements' consensuses multi-FASTA file as described above, a server that hosts a Ruby on Rails web application that acts as the interface with the user, and a high-performance server with the bioinformatics scripts (written in PHP) which, when invoked by the web application, runs Megablast and returns the result. This result is then parsed by the web application and graphically displayed. On the high-performance computer server, a script, running at cron time, takes care of updating the database blast index, retrieving the sequences from the local file server repository.
The GMO-Scan application is currently deployed in the intranet of the JRC and is used both to assist screening of sequences of suspected GMO origin and to quickly annotate new incoming sequences to the EU-RL GMFF. It also allows an efficient screening of contigs produced by Next Generation Sequencing experiments, and may represent an important element of the analyses pipelines that are being developed for efficient GMO detection and sequence analysis. The logic and infrastructure behind the GMO-Scan application are presented, and a number of use-cases are described to illustrate its potential.

H25: Amos Folarin, Caroline Johnston, Zina Ibrahim, Mark Begale, David Mohr and Richard Dobson. Exploiting the Quantified Self for clinical care: a framework for integrating mobile and sensor data into the electronic health record

Abstract: The Biomedical Research Centre for Mental Health (BRC-MH) is a partnership between the Institute of Psychiatry, King’s College London and the South London and Maudsley NHS Foundation Trust (SLaM).
Rapid and continued advancement in 'omics technologies have provided us with a wealth of high resolution information on the inner workings of disease and genetic variability. A

collateral rise in technological advances and widespread availability of mobile phones has at the same time been observed. These mobile technologies promise the potential of gaining equally fine grained information on patient behaviour and environment, especially when paired with additional small wearable probes that report specific information on the user. The ability to capture this data and report it back into the clinical record will be of immediate benefit, such as providing clinical trials with high resolution data for patient stratification, real-time response to treatment, and directly improve patient care and efficiency of intervention.
SLaM employs an electronic health record (EHR): the Patient Journey System (PJS) and has implemented a de-identified derivative of this EHR for research use: the Case Register Interactive Search (CRIS). Research projects can apply for access to CRIS data through a well-developed governance framework. To give patients a more interactive relationship with their clinical record, SLaM is using MyHealthLocker, a system built on top of Microsoft HealthVault. Clinicians can elect to share EHR data with their patients and patients can submit data to their own health record and elect to share parts of that record with their treatment team.
The Purple Robot application, developed by the Centre for Behavioural Intervention Technology (CBITS) at Northwestern University, provides both an interface to the on-board mobile phone probes (e.g. gps, light, temperature) and also a framework for integrating other mobile-aware probes e.g. worn accelerometers, oxygen saturation, heart rate, blood pressure, blood glucose etc. Using this platform, we have built a framework to enable patient reported outcome (PROM) data from smartphone applications, smartphone sensors and other sensor devices to be collected and analysed. Pertinent summary data can then be exported into HealthVault and, through a MyHealthLocker widget, patients can access and visualise their data and choose to share that data with their care team or researchers.
We have developed a proof-of-concept application which tracks sleep (quantity and quality) using mobile phone embedded sensors (GPS, light), a secondary wrist-worn accelerometer sensor (Fitbit) and patient reported information provided through an Android application. Data are cached in HealthVault and patient and researcher/clinician views presented through MyHealthLocker.

H26: Gabin Personeni, Simon Daget, Céline Bonnet, Philippe Jonveaux, Marie-Dominique Devignes, Malika Smail-Tabbone and Adrien Coulet. Mining Linked Open Data : a Case Study with Genes Responsible for Intellectual Disability

Abstract: Thanks to recent initiatives in the Linked Open Data (LOD) community (for instance, the Bio2RDF and EMBL-EBI RDF platforms), numerous life sciences resources have been made available on SPARQL endpoints, allowing for simplified programmatic access to these resources. We present an approach allowing for querying and integrating Linked Open Data on genes, with regard to an expert-defined Entity-Relationship (ER) model. We then show how we apply Inductive Logic Programming (ILP) to this data, in order to characterize genes responsible for intellectual deficiency (ID).
The first step of our approach is to build the ER model describing and typing the data to be collected: entity types and relationships of the model are mapped to RDF classes and properties, or description logic constructors of classes or properties. On top of that, identity conditions are defined, for each entity type, on instances of that entity type. There are three main types of conditions:
-RDF property that states and equivalence between two instances, such as owl:sameAs.
-Attributes of some instances, such as the HGNC symbol of a gene.
-Embedded data in URIs. For instance, we can deduce from the URIs of the gene <http://bio2rdf.org/geneid:5091> and the gene

<http://bio2rdf.org/kegg_vocabulary:hsa:5091> that they are identical, as they share the GeneID 5091.
Given an initial list of genes, the ER model mapping to the LOD allows to systematically build SPARQL queries to collect the data. Also the identity conditions are used to automatically "merge" equivalent instances in our local data model. In our study, we have collected data such as protein interactions, reactions, pathways and Gene Ontology (GO) annotations.
Inductive Logic Programming (ILP) allows to learn from the collected relational data, in order to characterize positive examples: in our case, these are genes responsible for ID identified in [Inlow, J. K. and L. L. Restifo, Genetics 166(2), 835-881(2004)] Characterization is done through theories, that are sets of First-Order Logic rules. Each rule describes a condition for a gene to be responsible. We ran four experiments (evaluated using leave-one-out cross-validation), each allowing a different degree of generalization w.r.t. GO term hierarchy, and an additional experiment without GO annotation facts.
The lack of GO annotation in the fifth experiment results in a low sensibility of the generated theory (26.6% of the positive examples are covered), but a high specificity (94.4% of negative examples are classified as negative). Comparatively, the best experiment using GO terms yields a theory covering 57.1% of the positives examples, at the cost of specificity (83.1%). Interestingly, theories with and without GO-terms annotations are very different. This suggests that they have complementary roled in characterizing genes responsible for ID.

H27: Christophe Becavin, Nina Sesto, Jeffrey Mellin, Francis Impens and Pascale Cossart. Listeriomics: Systems biology of the model pathogen Listeria

Abstract: Over the past three decades Listeria has become a model organism for host-pathogen interactions, leading to critical discoveries in a broad range of fields including virulence-factor regulation, cell biology, and bacterial pathophysiology. More recently, the number of Listeria “omics” data produced has increased exponentially, not only in term of number, but also in term of heterogeneity of data. There are now more than 40 published Listeria genomes, around 400 different transcriptomics data and 10 proteomics studies available. The capacity to analyze these data through a systems biology approach and generate tools for biologists to analyze these data themselves is a challenge for bioinformaticians.
To tackle these challenges we have developed a web-based platform named Listeriomics which integrates different type of tools for “omics” data manipulation, the two most important being: 1) a genome viewer for displaying gene expression array, tiling array, and RNASeq data along with proteomics and genomics data. 2) An expression atlas, which is a query based tool which connects every genomics elements (genes, smallRNAs, antisenseRNAs) to the most relevant “omics” data.
Our platform integrates already all genomics, and transcriptomics data ever published on Listeria and will thus allow biologists to analyze dynamically all these data, and bioinformaticians to have a central database for network analysis. Finally, it has been used already several times in our laboratory for different types of studies, including transcriptomics analysis in different biological conditions, and whole genome analysis of Listeria proteins N-termini.

H29: Vladimir Gligorijevic, Vuk Janjic and Natasa Przulj. Integration of molecular network data reconstructs Gene Ontology

Abstract: Gene Ontology (GO) describes genes and gene products in terms of their associated biological process (BP), molecular function (MF) and cellular component (CC).

GO was manually curated by domain experts and members of the research and annotation communities. However, due to the inconsistency in translation to GO terms and relations, manual curation has encountered many difficulties. Additionally, rapid development of technologies for biological data acquisition has resulted in an accumulation of biological data exceeding our ability to interpret them. To overcome
these problems, many computational tools for automatic gene and protein annotation have been devised. Recently, a shift was made from using GO to evaluate data to using data to construct and evaluate GO [1]. This recent work provides evidence that a large part of GO can be reconstructed from topologies of molecular networks. Dutkowski et al [1]. incorporate molecular interaction networks into the probabilistic clustering procedure to reconstruct GO and to identify new terms and relations that are missing from GO. Motivated by this work, we developed a novel data integration method which takes multiple types of molecular network data and utilizes them to reconstruct and update GO with new information. The method is based on Penalized Non-negative Matrix Tri-Factorization for heterogeneous data clustering. It takes all network data in a matrix form and performs simultaneous clustering of genes and GO terms inducing new associations (annotations) between genes and GO terms and between GO terms themselves. To improve the accuracy of our predictions, we extend the integration methodology to include additional topological knowledge represented as the similarity in wiring around non-interacting genes. We measure this by graphlet distance vectors (GDV) [2]. To make our study directly comparable to the competing method for reconstructing GO from network data, we apply our method on the same Saccharomyces cerevisiae network data as Dutkowski et al.[1].
Surprisingly, by integrating topologies of bakers yeast's protein-protein interaction, genetic interaction and co-expression networks, our method reports as related 96% of GO terms that are directly related in GO by "is_a", "part_of", "regulates" and "has_part" relationships. Our inclusion of GDV similarity of non-interacting genes contributes 6% to this large GO-term association capture. In
addition, our integration method outperforms previous methods by reproducing 48% of CC, 41% of MF, and 41% of BP GO terms. Furthermore, we use our method to infer new relationships between GO terms solely from the topologies of these networks and validate 44% of our predictions in the literature. Finally, we predict GO annotations of yet unannotated yeast genes and validate our predictions through genetic interactions profiling.[1] Dutkowski, J. et al., Nat Biotech, 31(1), 38–45, 2013 [2] Przulj, N., Bioinformatics, 23(2), e177–e183, 2007

H30: Matthias Ziehm, Aditi Bhat, Dobril K. Ivanov, Matthew D. Piper, Linda Partridge and Janet M. Thornton. Computational Biology of Ageing and Longevity in Model Organisms – Meta-Analyses of Survival Data and the SurvCurv Online Resource

Abstract: Understanding the biology of ageing is highly desirable because of potential benefits for a wide range of ageing-related diseases. However, it is also extremely challenging because of the underlying complexity. Survival records of longevity experiments are a key component in research on ageing, since ageing alone determines the mortality, if external hazards are excluded. A multitude of factors, including environmental, chemical and genetic ones, have been examined for their effect on longevity in various organisms from yeast to mice in numerous sometimes yearlong experiments. To make the most of these over-arching phenotype data, there is a need to collect them and provide a simple to use analysis platform allowing to compare and contrast the experiments and jointly analyse compatible data. This is even more important with the recent developments of automated high-throughput lifespan experimental procedures for invertebrate model organisms, especially Caenorhabditis elegans, promising an upcoming wealth of lifespan data.

Here we present SurvCurv, the first publicly available database and online resource for survival data in model organisms. All data in SurvCurv are manually curated and annotated. The database, available at www.ebi.ac.uk/thornton-srv/databases/SurvCurv/, offers various functions including plotting, Cox proportional hazards analysis, mathematical mortality models as well as various statistical tests. This allows users to reanalyse public data, analyse their own data and compare it with the largest repository of model-organism data from published experiments, thus unlocking the potential of survival data and demographics in model organisms. SurvCurv also facilitates method development for survival data by making these data available and enables the introduction of historical controls to research on ageing, a concept already established for cancerogenicity and toxicity in rodents.
We used the collected large set of survival data to address questions which are beyond the scope of individual studies. We performed multiple meta-analyses on Drosophila examining sex-differences and the potential dependence of lifespan extension effect on the survival times of the controls. We showed that there is no evidence for such dependence in Drosophila, indicating that lifespan extending treatments are not purely rescuing weak backgrounds. We further analysed the variation in survival of control cohorts under highly similar conditions and revealed that transgenic constructs of the UAS/GAL4 expression system, which should have no effect, extend lifespan significantly in the Drosophila strain w1118. Finally, we compared different model organisms in terms of the baseline survival characteristics across multiple experiments to determine the similarity of the ageing process phenotypically.

H31: Ranjeet Bhamber, Yan Zhang, Isabel Riba-Garcia, Julian Selley, Richard Unwin and Andrew Dowsey. The seaMass Framework: Peptide feature extraction from raw mass spectrometry data with non-negative sparse Poisson regression and learnt predictors

Abstract: Liquid Chromatography interfaced to Mass Spectrometry (LC-MS) has emerged as a central technique for elucidating patterns of protein changes between groups of biological samples. The output of LC-MS is a two-dimensional histogram constructed by binning each detected molecule by its LC retention time in the first dimension and MS mass-to-charge ratio (m/z) in the second. While LC is low-resolution, the MS instruments used are often highly reproducible and high resolution. A typical output has 5,000×100,000 bins or greater, resulting in dataset sizes >2Gb. While the number of unique peptides present in the sample is sparse in comparison, the raw data is invariably dense because: (a) peptides exhibit patterns of isotope peaks; (b) each peak is spread over multiple bins; (c) significant periodic background signal is evident. Downstream protein reconstruction, differential analysis and pathway modelling is dependent on peptide feature extraction from a preceding data processing phase. A typical software pipeline includes self-contained deterministic steps of Gaussian noise reduction, baseline subtraction, peak picking, and peak clustering. With this workflow errors propagate, overlapping signals are disregarded and evidence for weak biological signals routinely missed.
We have developed a framework that combines all aspects of the conventional workflow into an integrated sparse regression model. As low counts dominate other noise, we employ a Poisson likelihood. Since we directly interpret regression weights as peptide abundances, they must be non-negative. Moreover, the mean response for each raw datapoint depends additively on the predictors. The problem is similar to that in astronomical science and PET imaging, so we adopted a recent L1-sparse formulation of the seminal Richardson-Lucy/EM algorithm. The non-negativity constraint also enabled us to design highly specific predictors. Non-negative matrix factorisation with minimum-volume simplex constraint was employed to learn a group of predictors for each m/z bin that summarise the population of peptide isotope peak patterns from a public protein database. Through this, non-negative combinations of predictors closely approximate genuine patterns whilst encoding other patterns poorly. These

learnt predictors and generic background patterns were combined with multiscale B-spline basis functions to also model variable peak shape.
The resultant feature extraction method is unique in that it naturally models, separates and accurately quantifies overlapping peptide signals. Against the state of the art we demonstrate marked improvements in detection and quantification, especially on weak and overlapping peptides. For example, on simulated data with ground truth, at 90% sensitivity our approach has an FDR of 8% compared to 50% of NITPICK, while for quantification reliability our approach had an R squared of 0.973 against 0.657.

H32: Hanqing Liao, Isabel Riba-Garcia, Richard Unwin, Jeffrey Morris and Andrew Dowsey. Group-wise Image Registration-nOrmalisation (GIRO): Retention time alignment and abundance normalisation for LC-MS data

Abstract: Retention time alignment and abundance normalisation is crucial in the analysis of differential expression using Liquid Chromatography-Mass Spectrometry (LC-MS). The proposed GIRO algorithm takes the profile-mode output signals of each sample from LC-MS as a 2D image. By applying group-wise image registration methodology the retention time variation of the same peptide features in different samples can be corrected [1]. Image registration is a well-established technique in medical image analysis, but has not yet been widely applied for proteomics or metabolomics. The crucial novelty compared to previous approaches is that GIRO utilises the intensity information directly rather than indirectly through feature detection, so alignment is not hampered by detection inconsistencies across samples. In GIRO, a novel L1-regulated B-spline deformation field is employed to estimate and correct retention time variation. Moreover, in order to cope with inhomogeneity in the ion current between images, during the registration the images are simultaneously fitted by a carefully designed B-spline field to normalise out both systematic effects and differential expression.
GIRO can be used as a pre-processing tool for conventional feature extraction-based workflows to reduce misalignment and improve inter-sample normalisation. It can also be used in our recently proposed new image based workflow [2]. By avoiding feature extraction in the early stages, more information from the samples is preserved. This allows sophisticated image-based statistical modelling to be applied, resulting in more sensitive protein differential expression analysis. In this workflow, the LC-MS output for each sample is first transformed into an image; these are then registered to account for chromatography inconsistency; and finally, a functional mixed-effects model is applied [3]. Since perturbations in the raw data are preserved, we are able to quantify FDR-controlled differential effects on large numbers of signals missed by conventional software approaches, as demonstrated against Progenesis LC-MS (Nonlinear Dynamics, Newcastle, UK).
[1] AW Dowsey, MJ Dunn, GZ Yang, Automated image alignment for 2-D gel electrophoresis in a high-throughput proteomics pipeline, Bioinformatics 24, 950-957, 2008.
[2] H Liao, E Moschidis, I Riba-Garcia, Y Zhang, RD Unwin, JS Morris, J Graham, A W Dowsey, A new paradigm for clinical mass spectrometry analysis based on biomedical image computing principles, Proceedings of the IEEE International Symposium on Biomedical Imaging, 2014.
[3] JS Morris, V B Baladandayuthapani, R C Herrick, P Sanna, H B Gutstein, Automated Analysis of Quantitative Image Data Using Isomorphic Functional Mixed Models with Applications to Proteomics Data, Annals of Applied Statistics 5, 894–923, 2011.

H33: Nina Verstraete, Ignacio Sanchez and Diego Ferreiro. Spatial organization and distribution of linear motifs in the Ankyrin repeat protein family and its binding partners

Abstract: Background: Interactions between proteins regulate cellular physiology. Many of these interactions involve the recognition of short peptidic regions (i.e. short linear motifs, SLiMs) which can be characterized by simple sequence patterns, usually found in intrinsically disordered regions or in loops connecting globular or transmembrane domains. These peptide-domain interactions are typically transient and often involve folding upon binding, challenging the lock-and-key paradigm of protein recognition.
Ankyrin-repeats domains are one of the most frequently observed protein-protein interactors in nature. These domains are composed of tandem arrays of recurrent amino acids that cooperatively fold into elongated structures that mediate molecular recognition with high specificity. Many ankyrin-binding sites are either predicted or demonstrated to correspond to extended peptides mimicking SLiMs.
Description : We present here an exhaustive analysis of linear motif identification in Ankyrin proteins and their binding partners. We searched for enriched or depleted SLiMs with respect to a random exploration of the sequence-space in the Ankyrin protein family and their partners. We also analyzed the spatial distribution of SLiMs along the protein sequences and describe how particular SLiMs are specifically distributed in the Ankyrin-containing proteins.
Conclusions : This computational work presents sequence and structure-based approaches to analyze linear motif-mediated protein interactions in the Ankyrin repeat protein family. We discuss that the presence of functional constraints can conflict with the Ankyrin-repeats domains folding dynamics which in turn modulate the evolution of biological interactions.

H34: Yasuhito Inoue and Yasutaka Nakata. Statistical Analysis of Adrenergic Receptors

Abstract: The adrenergic receptors (ARs) are a family of G-protein-coupled receptors (GPCRs) that are targets for drug design in pharmacology field. There are three subfamilies of ARs, alpha 1, alpha 2 and beta receptors based on coupling to G-proteins. Several research groups have reported methods for the recognition and classification of the GPCRs. Although their methods are useful, there are some problems in applying these methods to a large protein family by the homology searches. Inoue et al. (2004) proposed that the transmembrane binary topology pattern is well conserved among GPCRs that have the same function, even though the amino acid sequences are not conserved, suggesting a possibility of classifying/identifying ARs based on the loop lengths contains some biological informations. In this study, we propose a novel classification method for classification of ARs using statistical analysis.
ARs with known transmembrane topology were extracted from UNI-PROT/SWISS-PROT (Release 56.0) (Bairoch et al. 2014) based on the following criteria: no description of "Fragment" or "Fragments" in the DE line, description of seven FT lines with "TRANSMEM". After removing the signal peptide region detected by SignalP 2.0 (Bendtsen et al. 2004), the transmembrane topology of these adrenergic receptor was predicted by applying HMMTOP 2.0 (Tusnády and I. Simon 2002). The final dataset consists of adrenergic alpha 1 receptor (17 sequences), alpha 2 (23), and beta (20) as defined by GPCRDB (Release 10.0; Horn et al. 2003) based on pharmacological properties and sequence similarities.
We calculated the fundamental statistical (minimum, mean, median, maximum, and standard deviation) values, and analyzed the loop lengths of ARs using by the classical multi-dimensional scaling of a data matrix (Gower 1966) using by R. Standard deviation values of N-terminus, intracellular 3rd, and C-terminus loops were larger than intracellular 1st and 2nd, ectracellular 1st and 2nd loops. For example, intracellular 3rd loop length of ARs was distributed from 47 to 179 a.a. around the mean (94.5 a.a.) and the median (72 a.a.). ARs were classified into three subfamilies using the classical multi-dimensional scaling method based on the distinctive loop lengths. Taken together, these findings indicated a possible functional and structural importance of loop lengths in the ARs.

H35: Stefan Naulaerts, Pieter Meysman, Wim Vanden Berghe, Bart Goethals and Kris Laukens. Mining the human proteome for conserved mechanisms

Abstract: All cells find themselves in continually changing environments to which they have to adapt, using their sensory system to provide input for the regulatory systems that integrate the information and trigger the eventual effectors. These cascades constitute a very complex cellular wiring that receives large amounts of attention from the scientific community due to its medical importance.
Recently, several great advancements in the understanding of these signaling cascades have been propelled by the popularity of "omics" approaches. The omni-present application of high-throughput analysis techniques has resulted in an unprecedented level of detail about gene expression and various aspects of cellular proteins, such as abundance, function and localization. In addition to this, vast public compendia of binary protein interactions for several model organisms have been created.
Although these information-rich inventories exist, the adaptive nature of protein complexes and signalling cascades remain poorly understood, as the current predominant mass spectrometric approaches are not suited to investigate the dynamic relationships between proteins. For example, binary protein interactions do not necessarily occur in vivo as the proteins could be expressed in different compartments of the cell or at different time points. This severely complicates the analysis of any protein interaction data. It thus remains a challenge to find out how biological entities cooperate to regulate cellular response to stimuli. Computational approaches can play a key role in the task of identifying and modelling signalling networks. Here, we present an integrative method, reliant on advanced pattern mining approaches to gain a deeper understanding of protein network dynamics. To this end, we created a compendium consisting of a large amount of proteomics papers for Homo sapiens that report differentially expressed proteins in cell lines. Next, we analysed this collection with frequent itemset mining to identify proteins that are often co-occurring in publications and used these patterns as the backbone structure of our further analysis. These patterns of co-occurring proteins were then enriched with additional attributes, such as gene expression correlation, protein localization and functional coherence metrics derived from the Gene Ontology tree and used as a filter on top of an integrated binary protein interaction network, obtained by fusing several of the most popular resources.
As a case study, we compared patterns derived from gene and protein expression data for several cancer types and cancer states (pre- and post-metastasis), which resulted in distinctly different patterns. We found that pattern-based analysis of the cascade of up- and downregulation on multiple "omics" levels can help to identify the cellular logic circuits and holds many promising applications in the biotechnological and biomedical areas.

H36: Pooya Zakeri, Leon-Charles Tranchevent and Yves Moreau. Kernel-Based Gene Prioritization Using Late Integration versus Geometric Kernel Fusion

Abstract: In biology, there is often the need to discover the most promising genes among large list of candidate genes to further investigate. In the recent years, several computational approaches based on different genomic data sources and often machine learning methods have been used to crack this problem efficiently. While a single data source might not be effective enough, fusing several complementary genomic data source results in more accurate prediction. Finding an efficient and cost-effective technique for merging these complementary genomic data has received increasing attention.
We consider several genomic data sources including annotation-based data sources, expression data sources, protein-protein interaction networks and literature, as well as sequence-based protein features such as information extracted directly from position-specific

scoring matrices and local sequence alignment.
Integrating data sources at the decision level, such as in the ensemble learning framework, is considered as an intuitive manner to deal with heterogeneous data sources. We discuss the advantage of modeling multi-heterogeneous biological data fusion in the prioritization task based on the ordered weighting averaging (OWA) operator and robust rank aggregation (RRA). We design several kernel-based gene prioritization frameworks that integrate multiple genomic data sources through late integration. After transforming genomic data matrices into kernel matrices, one kernel-based ranking approach is trained with known disease genes on each data type. Then fusion strategies defined by the use of OWA operator weights and ordered statistics weights are employed to aggregate performance rankings.
Moreover, we propose a kernel-based gene prioritization framework through genomic data fusion which is our kernel-based data fusion framework, recently developed and considered as a powerful tool for protein fold classification. It has been observed that geometric data fusion is less sensitive in dealing with complementary and noisy kernel matrices compared to typical multiple kernel learning approaches. Since genomic kernels often encodes the complementary characteristics of biological data, this motivates us to see the application of geometric data fusion in the gene prioritization task.
Finally, we evaluate our models by applying them to the disease data sets on which ENDEVAOUR has been benchmarked. Our late integration frameworks based on OWA or RRA offers an improvement in empirical performance. In addition, it is observed that using geometric kernel fusion, we achieve competitive results compared with best existing MKL approaches proposed for gene prioritization task.

H37: Xin He, Ernest Walzel, Douglas Armstrong and Ian Simpson. DisEnT, a unified ontology based gene set enrichment analysis (GSEA) framework for gene disease and gene phenotype studies.

Abstract: Researchers often want to derive biological insight from lists of genes that have emerged from experimental and computational studies. A common approach is to perform gene set enrichment analysis (GSEA) on such lists, typically using ontology based gene-association datasets such as those generated by the Gene Ontology (GO) Consortium that annotate cell component, molecular function and biological process terms to genes. The objective is to establish whether any particular biological properties are annotated to genes of the list in a higher than expected number. The effectiveness of any such approach is dependent on the quality and the coverage of the gene-association data across species. Whilst iterative improvements to GO association data have been ongoing for decades, many other established and emerging ontologies that would be beneficial for biological interpretation of gene lists have poor and inconsistent coverage.
In order to facilitate the adoption of other ontologies and data corpora for ontology based GSEA we have developed an integrated and generic framework that first uses NCBO-Annotator and MetaMap to map ontology terms from free text, then integrates this multi-sourced mapping data into a unified relational database system for GSEA analysis. Finally we use a modified version of the R/Bioconductor TopGO package to calculate topology aware statistics for term annotations between source and background gene lists.
We demonstrate our workflow by processing disease data from three publicly available data sources including OMIM, PubMed-GeneRIF and EBI-VARIATION. We present a comprehensive and integrated gene to disease association dataset annotated to the human disease ontology (HDO) that we call the Disease Enrichment Tool (DisEnT) framework. We show the utility of the framework, including validation approaches with several use-case examples for neuropsychiatric diseases based on lists of proteins from the pre- and post-synaptic compartments and of genes implicated in learning and memory processes through

high-throughput gene expression assays for behaviour. Finally we present an easy to use web application for using the DisEnT system for general use.

H38: Lucila Aimo, Robin Liechti, Anne Niknejad, Nevila Nouspikel, Anne Gleizes, Dmitry Kuznetsov, Fabrice David, Vassily Hatzimanikatis, Howard Riezman, F. Gisou van der Goot, Lydie Bougueleret, Ioannis Xenarios and Alan Bridge. SwissLipids – a knowledge resource for lipid biology

Abstract: Lipids are an important and diverse class of molecules. They form the bulk of cellular membranes, constitute the primary means of long-term cellular energy storage, and play important roles as signaling and regulatory molecules. According to recent estimates, the lipid complement or 'lipidome' of humans may consist of tens thousands of distinct lipids. To understand how this complexity is generated and how it affects cellular and organismal function, the Swiss Initiative for systems biology SystemsX.ch launched the project LipidX (http://www.lipidx.org/), which uses a high throughput mass spectrometry-based experimental platform to survey thousands of defined lipid species in a variety of experimental systems. To maximize the value of the data generated by LipidX and to facilitate its integration with prior biological knowledge we have developed an expert-curated knowledge resource on lipids and their biology – SwissLipids. SwissLipids is built using information on lipid metabolism which is first curated in the ChEBI ontology of chemical entities (http://www.ebi.ac.uk/chebi/) and the Rhea database of biochemical reactions (http://www.ebi.ac.uk/rhea/). This curated information provides the building blocks – lipid classes, fatty acids and alcohols – for the enumeration of a set of over 150,000 theoretically feasible glycerophospholipid and glycerolipid structural isomers. These isomers are annotated with standard names (following LIPID MAPS conventions), cheminformatics descriptors (InChI, SMILES, mass, formula), and mappings to the ChEBI namespace, the metabolic reactions of Rhea, and other databases (such as LIPID MAPS). They are arranged in a hierarchical classification which reflects the degree of structural information provided by common analytical methods such as (tandem) mass spectrometry (J. Lipid Res. 2013, 54(6):1523-30), generalizing from structural isomers – with a complete specification of stereochemistry and double bond position, and orientation – to lipid species, with information on lipid class and mass. The hierarchy includes over 200,000 lipids (including isomers), and is fully mapped to the ChEBI ontology and LIPID MAPS classification at all levels. It serves as a framework for the expert curation of published information on the biological roles and interactions of lipids using well supported community standard vocabularies and ontologies such as UniProt identifiers for proteins, Gene Ontology (GO) terms for functions, processes, and locations, UBERON for tissues, and the evidence code ontology (ECO). The current version of the lipid hierarchy includes curated links to published knowledge from over 270 peer-reviewed papers. We are continually updating the lipid hierarchy with new lipid classes and new expert curated knowledge, and plan to make this information available through a dedicated web-based interface at http://swisslipids-dev.vital-it.ch/ in late 2014.

H39: Nisar Shar, Vijayabaskar Ms and David Westhead. Predicting transcription factor mutual interactions from ENCODE data

Abstract: Transcription factors play an important role in the regulation of gene expression. They bind to genomic DNA, in promoter or enhancer regions, and influence transcription. Although the mechanism of influence is not clear in every case, it may include direct interactions with the transcriptional machinery and the recruitment of accessory molecules such as chromatin modifiers. The ENCODE project has used the ChIP-seq technique to map the binding sites of 119 transcription factors out of more than 1800 human transcription

factors in 72 cell types, along with information on DNase I hypersensitivity and chromatin modification. It is now widely accepted that genetic regulation is combinatorial in nature and recent studies show that mutual interactions of TFs can be predicted if ChIP-seq binding profiles of two TFs overlap each other. We have developed and evaluated statistical methods for assessing significantly overlapping binding in between TF pairs genome wide. We predicted the known protein-protein interactions and also some novel interactions in one cell line (Gm12878) by using ChiP-seq data and these interactions were evaluated by the statistical methods.

H40: Jian-Long Huang, Ming-Yi Hong and Chien-Yu Chen. SeqHouse: a web platform for biological data integration and management

Abstract: With the great advance of next-generation sequencing (NGS) technologies, NGS data analysis and interpretation have become more and more common in bioinformatics research. A lot of computational tools and web services, such as sequence assemblers, sequence alignment tools, and genome mappers, have been developed and widely used for genome-wise studies. Current efforts on integration and organization of the analyzed results from public NGS data sets for further queries and statistics have been made, but some challenges have not been fully addressed and resolved. For example, most of the existing data warehouses, such as InterMine, BioMart and Atlas, do not allow the users to upload analyzed results through web interface. In other words, the abovementioned data warehouses usually require administrators for data uploading and management. In this regard, a web-based platform named SeqHouse is proposed here for effectively and efficiently managing the analyzed results of NGS data. It is designed and constructed for researchers with biological background to conveniently upload, annotate, query, and export the analyzed data, such as assembled sequences, sequence alignments, genome mapping results, and gene expression profiles. SeqHouse accepts commonly used data formats, such as FASTA, GFF, and tab-delimited files, and converts them into database objects for indexing and creating the object relationships, which can be used for invoking further analysis pipelines to produce more useful results efficiently. For example, SeqHouse can automatically identify the orthology group for each transcript based on the results of blastx against the nr database; mutation detection can be performed on assembled transcripts from different strains; and differential analysis can be performed on the uploaded expression data from RNA-seq, microarray, or qPCR experiments of two conditions. SeqHouse is a user-oriented platform, in which every user can register in order to upload their own analyzed results and decide which data can be shared with other users for queries. The eventual goal of the SeqHouse is not only to provide the interface for data storage and management, but also an easy-to-use environment for mining and discovering valuable information.

H41: Mauro Petrillo, Alexandre Angers, Alex Patak and Joachim Kreysa. Screening of public nucleotide databases with PCR simulation to generate a secondary database containing sequences related to Genetically Modified Organisms.

Abstract: The Joint Research Centre (JRC) of the European Commission is host to the European Union Reference Laboratory for Genetically Modified Food and Feed (EU-RL GMFF). The core tasks of the EU-RL GMFF are the scientific assessment and validation of detection methods for GM Food and Feed as part of the EU authorisation procedure.
The EU-RL GMFF maintains a database of Polymerase Chain Reaction (PCR) - based reference methods for GMO analysis, called “GMOMETHODS”, which is publicly available on its website. In order to identify sequences potentially linked to GMOs and to evaluate the specificity of the reference methods while also providing information on their potential

amplicons, we created the JRC GMO-MethodScan-DB, a nucleotide database of GMO-related sequences, obtained by screening of public nucleotide sequence databanks, including patents and available whole plant genomes, with a PCR simulation tool.
The platform behind the JRC GMO-MethodScan-DB is built of three servers: a server dedicated to the database (PostgreSQL) that stores all the DNA sequences and related information, a server that hosts a Ruby on Rails web application that acts as the interface with the user, and a high-performance computer that regularly runs a compute-intensive pipeline. Given a database of nucleic acid sequences and the detection methods of the GMOMETHODS database, the pipeline performs the following tasks: A. in silico PCR simulation on the whole sequence set with the whole detection method set (allowing for gaps and mismatches) by running e-PCR, a tool developed at NCBI and installed locally; B. Retrieval of the matching sequences and extraction from them of the sub-regions (amplicons) identified by e-PCR; C. if necessary, alignment of probe with the amplicon sequence by running matcher, a sequence comparison tool of the EMBOSS package; D. if possible, species assignment to the amplicon sequences; E. generation of the output as TAB delimited file with all metadata, then used to populate/update the PostgreSQL database. Currently the pipeline runs on the EBI non-redundant patent nucleotide database NRNL1, the NCBI nt collection and the ENSEMBL whole plant genomic sequences.
The JRC GMO-MethodScan-DB is now deployed as testing phase in the intranet of the JRC and is used both to assist the validation of new GMO detection methods, in particular when unexpected amplifications or PCR reactions with lower performances than expected are observed in the laboratory, and to verify the specificity of already validated methods, as new sequence information is continuously made available and needs to be assessed. Moreover, we are fishing published sequences that were not reported as related to GMO events, in particular when screening patent sequence sets. In the near future, we are planning to make the tool publicly available to the research community, to assist laboratories in the assessment of the GMO detection methods that they use.

H42: Aida Mrzic, Trung Nghia Vu, Dirk Valkenborg, Evelyne Maes, Filip Lemière, Bart Goethals and Kris Laukens. Pattern detection in associated artifact peaks in mass spectra with frequent itemset mining

Abstract: Even high quality mass spectra that lead to confident protein identifications often contain a significant number of unexplained masses. These masses are typically not further analysed, but are often assumed to be contaminants, modified ions, or non-protein components in the sample. Software tools either use these unmatched masses to attempt to discover unexpected modifications or to perform a second round search, or eliminate aberrant mass masses in a preprocessing step. Most investigations concerning the origin of aberrant spectral masses take into account the occurrence of single ion masses. They do not use or investigate the fact that frequently co-occurring unassigned peaks are likely to have a common origin.
We introduce a framework to discover patterns of frequently co-occurring aberrant peaks in unexplained historical mass spectrometry data. We propose unsupervised pattern mining techniques to reveal which peaks are associated, and thus are likely to have a common origin. Our pattern mining approach is based on a hypothetical spectral model, in which observed ions are classified according to their origins. The technique to discover associations between frequently co-occurring peaks is frequent itemset mining, a class of pattern detection techniques that is specifically designed to discover co-occurring items in transactional datasets. The archetypical example of frequent itemset mining is the discovery of products that are frequently purchased together from mining large numbers of supermarket basket transactions.

We evaluated our approach on both MS and MSMS datasets. The first case study was a dataset consisting of 443 MALDI TOF spectra from gel separated and tryptically digested proteins. After peak extraction and filtering, the peak lists were compared to the relevant SwissProt sequence databases using Mascot. All assigned peaks from spectra that yielded positive hits in Mascotwere eliminated. From the remaining spectra, all unassigned peaks were extracted, and discretized to their nominal masses. Analysis of the unexplained fraction of the dataset reveals several highly frequent masses, and pattern mining demonstrates that many of these peaks frequently co-occur. Several of these co-occurring peaks indeed were shown to have a known common origin, whereas other patterns are promising hypothesis generators for further analysis. The approach was also extended to an MSMS dataset, in order to find patterns in the mass differences between unexplained peaks of a single spectrum. Bot trivial and less trivial patterns were obtained and are currently being interpreted. The method can be applied to historical laboratory data to generate interesting patterns that emerged throughout history. For new experimental data, such patterns can give valuable hints for interpretation, quality control and data dependent acquisition workflows.

H44: Vidya Oruganti, Martin C. Simon and Marcel H. Schulz. Small RNA Analysis and Visualization in P. tetraurelia

Abstract: Several independent mechanisms were described how RNA passes acquired genetic and epigenetic information to sexual progeny: (i) epigenetic inheritance by affecting the DNA composition of daughter cells, (ii) epigenetic control of DNA copy number of daughter cells, and (iii) epigenetic inheritance of acquired point mutations. Especially the recent work in Paramecium tetraurelia (P. tetraurelia) contributes to our understanding of RNA as a transgenerational carrier of information [1].
The prevalence of small RNAs across a range of species has tempted the development of computational tools to predict and analyse small RNA biogenesis and interaction. There has been little focus on the visualization of small RNA clusters. We present an automatic pipeline for visualization and analysis of small RNA reads that highlights, among other things, differences in strand occupation and non-templated small RNA modifications.
We showcase our pipeline on the analysis of small RNA sequencing conducted for different temperature environments and vegetative states in P. tetraurelia. We quantified and compared these datasets to determine their possible precursors, RNA structures and phased loci.
We find that the predominant length of small RNAs is found to be 23bps. We also find the first indication that phasing can act as a mechanism to produce sets of small RNAs; these phased loci are common between most of the datasets.
References
[1] Simon & Plattner, Unicellular eukaryotes as Models in Basic Cell and Molecular Biology - a critical appraisal of their past and future value , Int Rev Cell Mol Biol, 2014

H45: Albert Pallejà Caro, Sune Frankild, David Westergaard, Pope Moseley and Søren Brunak. Linking ICD10/SNOMED CT concepts to human genes

Abstract: Background: The healthcare sector extensively uses the WHO International Classification of Diseases (ICD)[1] to codify diseases when they diagnose patients in electronic and non-electronic patient health records. The aim of this work is to produce a mapping between the ICD version 10 and Ensembl gene identifiers. The mapping will facilitate carrying out systems biology analyses based on electronic patient records, with the aim of finding the underlying causes of disease and for example the molecular links between diseases and their comorbidities.
Methods: To map the ICD10 codes to proteins, we created an intermediate mapping between

ICD10 and the Disease Ontology (DO)[2]. To our knowledge, no mapping between the ICD10 and DO terms is publicly available. The DO is a standardized hierarchical ontology for human diseases. Both ICD10 and DO dictionaries are more phenotype-oriented, which makes it easier to make a reliable mapping between them. We included all the ICD10 codes at level 3 contained in chapters II-XV and XVII. The following two rules were applied when mapping ICD10 codes to equivalent DO terms: 1) if there was a direct or a meaningful match between an ICD10 code description and a DO term or one of its synonyms, we assigned the ICD10 code to that DO term; 2) if there was not a meaningful match between an ICD10 code and a DO term, we went up the DO hierarchy until we found a parent DO term that could include the ICD10 code and all its complexity. For the relationship between DO terms and human genes, we have extensive knowledge from both the Uniprot-KB[3], the Genetic Home Reference (GHR)[4] and an internal text-mining project. The methodology of the pipeline is described in ref. 5.

Results: We could manually map 1,050 ICD10 level 3 codes to 490 DO terms. When combining the previously mentioned resources, this resulted in 883 ICD10 level 3 codes, which have at least one gene associated. There are a total of 2,982 unique proteins, and 59,828 ICD10-protein associations in the mapping. The median number of genes per ICD10 code is 11. We find that the distribution of genes per disease is bimodal, reflecting that some ICD10 codes are very broad, and are mapped onto equivalently broad DO terms. Looking at the number of disease-gene relations by chapter, we find that medical research is dominated by cancer research, congenital malformations and endocrine disorders.

[1] WHO. International Classification of Diseases. (2010). http://apps.who.int/classifications/icd10/browse/2010/en

[2] Schriml, L. M. et al. Disease Ontology: a backbone for disease semantic integration. NAR. 40, D940–6 (2012).

[3] The UniProt Consortium. Activities at the Universal Protein Resource (UniProt). NAR. 42, D191–8 (2014).

[4] National Library of Medicine. Genetics Home Reference. (2013). http://ghr.nlm.nih.gov/

[5] Franceschini, A. et al. STRING v9.1: protein-protein interaction networks, with increased coverage and integration. NAR. 41, D808–15 (2013).