Biocuration 2016 - Posters

Source: http://www.sib.swiss/events/biocuration2016/posters

1 RAM: A standards-based database for extracting and analyzing disease-specified concepts from the multitude of biomedical resources

Jinmeng Jia and Tieliu Shi

Each year, millions of people around world suffer from the consequence of the misdiagnosis and ineffective treatment of various disease, especially those intractable diseases and rare diseases. Integration of various data related to human diseases help us not only for identifying drug targets, connecting genetic variations of phenotypes and understanding molecular pathways relevant to novel treatment, but also for coupling clinical care and biomedical researches. To this end, we built the Rare disease Annotation & Medicine (RAM) standards-based database which can provide reference to map and extract disease-specified information from multitude of biomedical resources such as free text articles in MEDLINE and Electronic Medical Records (EMRs). RAM integrates disease-specified concepts from ICD-9, ICD-10, SNOMED-CT and MeSH (http://www.nlm.nih.gov/mesh/MBrowser.html) extracted from the Unified Medical Language System (UMLS) based on the UMLS Concept Unique Identifiers for each Disease Term. We also integrated phenotypes from OMIM for each disease term, which link underlying mechanisms and clinical observation. Moreover, we used disease-manifestation (D-M) pairs from existing biomedical ontologies as prior knowledge to automatically recognize D-M-specific syntactic patterns from full text articles in MEDLINE. Considering that most of the record-based disease information in public databases are textual format, we extracted disease terms and their related biomedical descriptive phrases from Online Mendelian Inheritance in Man (OMIM), National Organization for Rare Disorders (NORD) and Orphanet using UMLS Thesaurus. Currently, RAM contains standardized 2,842 rare disease records, 27,329 phenotypes and 75,335 symptoms. Each record-based disease term in RAM now has 8 annotation fields containing definition, synonyms, symptom, phenotype, causes, diagnosis, treatment and cross-linkage, 5 of them have been presented in both textual and standards-based structural format. We continue updating RAM by mapping and extending standardized concepts for each annotation field's content through ICD, SNOMED-CT, Human Phenotype Ontology (HPO) and UMLS CUI. which make RAM as a standardized information system for biomedical information exchange and data integration between different standard-compliant databases.

2 A computational network model describing xenobiotic metabolism response in the liver built using the semi-automated curation workflow BELIEF.

Justyna Szostak, Iro Oikonomidi, Giuseppe Lo Sasso, Marja Talikka, Sam Ansari, Juliane Fluck, Sumit Madan, Florian Martin, Manuel-C. Peitsch and Julia Hoeng

Mechanistic comprehension of the impact of treatment on disease initiation and progression calls for techniques that convert ever-increasing literature-based scientific knowledge into a format that is suitable for modelling, reasoning, and data interpretation. The BEL Information Extraction workFlow (BELIEF) facilitates the transformation of unstructured information described in the literature into structured knowledge. BELIEF automatically extracts causal molecular relationships from text and encodes them in BEL statements. BEL (Biological Expression Language) is a standard language for representing, integrating, storing, and exchanging biological knowledge extracted from scientific literature in a computable form. The assembled BEL statements are used to build biological network models that can link the upstream events to downstream measureable and quantifiable entities, represented as differential gene expression.Using BELIEF, we have built a biological network model that describes xenobiotic metabolism in the liver context. Liver is a critical organ responsible for the elimination of toxic compounds by converting them into suitable forms during the xenobiotic metabolism process. During this process, the liver is susceptible to injury although the molecular aspects of this injury are not completely understood. Through the BELIEF platform, we were able to create a network model that captures important players in the context of liver xenobiotic metabolism. Nuclear receptors and transcription factors, such as the aryl hydrocarbon receptor, play a pivotal role in this comprehensive network model. This network model also includes the signalling pathways that lead to the activation of enzymes responsible for phase I (e.g., CYP1 family) that convert lipophilic chemical compounds into their hydrophilic forms. It also activates phase II conjugation enzymes (mainly the transferases), responsible for processes such as glucuronidation, sulfation, methylation, and acetylation as well as phase III membrane transporters, responsible for the elimination of xenobiotic metabolites.This two-layered causal biological network model may represent a step forward in toxicological risk assessment and improve our understanding of how different toxicants are metabolized by the liver.

3 QuickGO: a web-based tool for Gene Ontology browsing, interpretation and analysis.

Aleksandra Shypitsyna, Melanie Courtot, Elena Speretta, Alexander Holmes, Tony Sawford, Tony Wardell, Sangya Pundir, Xavier Watkins, Maria Martin and Claire O'Donovan

The Gene Ontology (GO) is a valuable resource for the functional annotation of gene products. In line with users' demands to accurately capture, interpret, analyze and evaluate the available functional information on genes and their attributes, its role in the scientific community is ever more important. GO is constantly expanding and improving to best represent our knowledge in various areas of biology, reflecting the development of new methods and the exponential growth of biosample resources and their experimental characterization. The current version of GO contains just under 45, 000 terms, to which over 250 million annotations have been made. It is therefore critical to be able to easily and quickly mine and visualise the available information. The Gene Ontology Annotation (GOA) team would like to announce the release of a new version of our popular web-based tool for browsing and interpreting the GO and all associated annotations, QuickGO. Benefiting from the constant communication with our international community (lab scientists, expert reviewers, bioinformaticians, clinicians and curators), the new QuickGO offers many more features in addition to improved speed, stability and data visualisation. Addressing the recent expansion of GO scope and aims, QuickGO now provides functional annotations not only to proteins, but also to protein complexes and RNA. The new user interface allows quick and easy searching of gene names, accessions, GO terms and annotations as well as evidence codes - recently connected to the Evidence Code Ontology. The interface, as proven by extensive user testing, is friendly and highly intuitive for users with different levels of GO experience. Additionally, it provides integrated contextual help covering all aspects of functionality. Following the needs and requests of the scientific community contributing to and employing GO data, the filtering options in QuickGO have been significantly redesigned. The basic and most popular filters can now be applied directly while visualizing annotations of interest, while experienced users have access to advanced filters to retrieve, interpret and further analyze their datasets.The front page of the QuickGO browser has direct links to major functions, including our GO slimming wizard. This tool allows mapping of the more GO granular terms to a smaller number of higher level, broader parent terms, and thus enables a quick overview of a genome or the large experimental datasets, of utmost importance for GO enrichment analysis. The wizard's design has been completely redeveloped to focus on simplicity and intuitive browsing, while preserving its accuracy and functionality. The release of this new version of QuickGO, powered by recent innovations in technology, reflects the development of GO and biological sciences. It addresses the various needs of our users from contributing to GO annotation to applying it to discover functional information about genes and gene products, interpreting the results from the particular biological disciplines or looking for new ways to direct research.

4 MSigDB 5.1: a database of human gene sets.

Arthur Liberzon, Helga Thorvaldsdottir, Pablo Tamayo and Jill P. Mesirov

Genetic screens, transcriptome profiling, and other kinds of genome-wide surveys report lists of genes as their main result. Analysis of this data is greatly facilitated by comparison to sets of genes categorized by properties they have in common. Gene sets define "what the genes do" by describing biological processes and results of genomic studies in a very simple and intuitive manner. For example, all genes known to function in a signaling pathway constitute a gene set. Gene Set Enrichment Analysis (GSEA) methods gain additional power by considering many genes as a group. The Molecular Signatures Database (MSigDB) was originally developed to supply gene sets for GSEA. A decade later, MSigDB is one of the most widely used and comprehensive repositories of gene sets. The MSigDB now contains over 13 thousand gene sets, which are organized into eight collections according to their derivation. The collections include genes grouped by chromosomal locations (C1), canonical pathways and lists curated from papers (C2), genes sharing cis-regulatory motifs (C3), clusters of genes co-expressed in microarray compendia (C4), genes grouped according to gene ontology associations (C5), as well as expression signatures of oncogenic pathway activation (C6) and immunology (C7). A special collection of "hallmark" gene sets (H) consists of signatures derived from multiple "founder" sets. The hallmarks summarize the original founder sets into signatures that display coherent expression in specific, clearly defined biological states.

5 The Practice of Medical Science Data Management and Sharing Platform (SDMSPM)

Zhang Ze, He Xiaolin, Sun Xiaokang, Qian Qing and Wu Sizhu

With the rapid development of computer science and information technology, a large amount of medical data are generated from medical research, medical experiments, medical services, health care, health management, etc. These data are essential for clinical diagnosis, research and hospital management. How to manage, store, organize, and use multiple types of data from various sources effectively has become an important challenge. Currently, multi-source data management and sharing are based on datasets, not data records, it is difficult to achieve a shared purpose. Therefore, according to medical data lifecycle, this study has explored to build a medical scientific data platform for efficient data sharing, management and utilization. We have designed and developed a platform, called Medical Science Data Management and Sharing Platform (SDMSPM). The architecture of SDMSPM contained four layers: a support layer, a data storage layer, a functional layer and a application layer, and supported by the data standards system and the security system. This platform uses the portal and B/S framework, focusing on unstructured data storage, data submission through web interface and batch mode, metadata specifications, data associationn, data classification and data sharing. It is worth mentioning that the system uses MongoDB to store unstructured data which uses BSON-format file and dynamic mode, so that a various types of data can be integrated easier and faster. In addition, we have realized the establishment of dataset metadata specification, data recording metadata specification and data sets classification criteria in datasets management and records management. We also have carried out semantic computing based on existing data fields or Medical Subject Headings (MESH) thesaurus to discovery more relationship between different datasets, enriching relationship storage and providing the associated services. Experimental data are from Population and Health Science Data Sharing Platform of China. We also have produced a special data analysis for the field of cancer, combining with the Google trends and PubMed statistical data. Html5, D3.js,Echart and other visualization techniques are applied for data dynamic display. User permissions and share permissions have also been set up to ensure the security of sharing platform.

6 The Type 2 Diabetes Knowledge Portal: accelerating type 2 diabetes research through community access to human genetic information and tools.

Maria Costanzo and Accelerating Medicines Partnership

The Type 2 Diabetes Knowledge Portal (http://www.type2diabetesgenetics.org) is an open-access resource for human genetic information on type 2 diabetes (T2D). It is a central repository for data from large genomic studies that identify DNA variants whose presence is linked to altered risk of having T2D or related traits such as high body mass index. Pinpointing these DNA variants, and the genes they affect, will spur novel insights about how T2D develops and suggest new potential targets for drugs or other therapies to treat T2D.The T2D Knowledge Portal aggregates data in a framework that facilitates analysis across disparate data sets while properly crediting researchers and protecting patient privacy. It provides a user-friendly interface that enables scientists to search for information on particular genes or variants and to build queries for the sets of variants associated with particular traits. Data will be added to the knowledgebase on an ongoing basis via the collection of existing data sets and the incorporation of new data sets as they become available, continually increasing the power of analyses that can be performed. Currently, data are stored in a Data Collection Center at the Broad Institute. In the near future, federated nodes at other sites will also receive data and will connect with the T2D Knowledge Portal to allow analyses across all data at all sites. Such federation will enable each node to protect individual patient data in accordance with local regulations while facilitating global access to analyses of the data.Financial support for this project is provided by the Accelerating Medicines Partnership in Type 2 Diabetes-a collaboration of the National Institutes of Health, five major pharmaceutical companies, and three large non-profits-and by the Carlos Slim Foundation.

7 Web-based 3D visualisation of anatomy at eMouseAtlas

Chris Armit, Bill Hill, Nick Burton, Lorna Richardson, Liz Graham, Yiya Yang and Richard Baldock

eMouseAtlas develops tools and resources that enable high-end visualisation of embryo data in the context of a web browser. Mouse embryo anatomy is delineated to a very high standard and is assigned EMAPA anatomy ontology terms, enabling queries across EMAGE and GXD gene expression databases. A section viewer allows visualisation of anatomy on arbitrary section through mouse embryos - much like a virtual microtome - whilst a 3D anatomy pop-up window allows users to visualise the delineated anatomical components in an interactive 3D-context as either a wireframe or surface-rendered model. The new viewer uses IIP3D and WebGL technology to allow interactive exploration of 3D anatomy in a HTML5-compatible and WebGL-enabled web-browser and without the need for data download.Both the WebGL navigation tool for the section viewer and the 3D anatomy pop-up window require surface generation of embryo models, and this requires curation effort and sub-sampling of 3D models. We report on our efforts to streamline this process to enable high-throughput visualisation of 3D embryo data. Furthermore, we report on our use of this 3D viewer in prototype web-based visualisation of 3D gene expression and phenotype data.URL: http://www.emouseatlas.org/eAtlasViewer_ema/application/ema/anatomy/EMA27.php

8 Mouse Tumor Biology (MTB) Database - Data Visualization via a Faceted Search Interface

Debra M. Krupke, Dale A. Begley, Steven B. Neuhauser, Joel E. Richardson, John P. Sundberg, Carol J. Bult and Janan T. Eppig

As the rate of data production continues to increase in the field of cancer research it is becoming increasingly important not only to capture that data but also to develop tools for its efficient searching and visualization. The Mouse Tumor Biology Database (MTB; http://tumor.informatics.jax.org) contains data on hundreds of tumor types, some having thousands of records. We have implemented a faceted, iterative search tool to facilitate efficient searching of large volumes of heterogeneous data to allow researchers to rapidly identify mouse models of human cancer most relevant to their research.   The faceted search page presents a summary of all data in MTB. As search terms are selected by the user, the corresponding search results are dynamically displayed. Subsequent search terms are limited to those relevant given the preceding search term selection and search results are updated automatically. Multiple terms may be combined to iteratively focus in on a small set of data. Results may be sorted by clicking the column header of the desired sort field. Color-coded tumor frequency summary information is provided to further aid in data review. Hyperlinks direct the user to more detailed data regarding each model.   Access to MTB is provided free of charge and presents researchers with tools to facilitate the identification of experimental models for cancer research. The data in MTB are primarily related to genetically engineered mouse models and Patient Derived Xenograft (PDX) models of human cancer. Data related to cancer cell lines are not comprehensively represented in MTB. Our expert biocurators use controlled vocabularies and internationally accepted gene nomenclature standards to aid the integration of data from a variety of sources. The primary source for data in MTB is the published literature supplemented by submissions from the cancer genetics research community.   MTB is supported by NCI grant CA089713.

9 Wikidata as a semantic framework for the Gene Wiki initiative

Sebastian Burgstaller-Muehlbacher, Andra Waagmeester, Elvira Mitraka, Julia Turner, Tim Putman, Justin Leong, Chinmay Naik, Paul Pavlidis, Lynn Schriml, Benjamin Good and Andrew Su

Open biological data are distributed over many resources making them challenging to integrate, to update and to disseminate quickly. Wikidata is a growing, open community database which can serve this purpose and also provides tight integration with Wikipedia.In order to improve the state of biological data, facilitate data management and dissemination, we imported all human and mouse genes, and all human and mouse proteins into Wikidata. In total, 59,721 human genes and 73,355 mouse genes have been imported from NCBI and 27,306 human proteins and 16,728 mouse proteins have been imported from the Swissprot subset of UniProt. As Wikidata is open and can be edited by anybody, our corpus of imported data serves as the starting point for integration of further data by scientists, the Wikidata community and citizen scientists alike. The first use case for these data is to populate Wikipedia Gene Wiki infoboxes directly from Wikidata with the data integrated above. This enables immediate updates of the Gene Wiki infoboxes as soon as the data in Wikidata are modified. Although Gene Wiki pages are currently only on the English language version of Wikipedia, the multilingual nature of Wikidata allows for usage of the data we imported in all 280 different language Wikipedias. Apart from the Gene Wiki infobox use case, a SPARQL endpoint and exporting functionality to several standard formats (e.g. JSON, XML) enable use of the data by scientists by both direct query and via mediating domain-specific applications.In summary, we created a fully open and extensible data resource for human and mouse molecular biology and biochemistry data. This resource enriches all the Wikipedias with structured information and serves as a new linking hub for the biological semantic web.

10 The BioGRID Interaction Database: Integration of Genetic, Protein and Chemical Interactions and an Improved Network Viewer

Rose Oughtred, Bobby-Joe Breitkreutz, Lorrie Boucher, Christie S. Chang, Jennifer M. Rust, Andrew Chatr-Aryamontri, Nadine Kolas, Lara O'Donnell, Chandra L. Theesfeld, Chris Stark, Kara Dolinski and Mike Tyers

Delineating protein and genetic interaction networks is critical to understanding complex biological pathways underlying both normal and diseased states. To further this understanding, the Biological General Repository for

Interaction Datasets (BioGRID) (www.thebiogrid.org) curates genetic and protein interactions for human and major model organisms, including yeast, worm, fly, and mouse. As of February 2016, BioGRID contains over 1,052,000 interactions manually curated from high throughput data sets and low throughput studies, as documented in more than 45,800 publications. This includes over 350,000 human interactions focused on normal and disease-related processes such as the Ubiquitin Proteasome System (UPS), which is implicated in metabolic, cardiovascular and neurodegenerative diseases, as well as cancer. Recently, BioGRID has begun to incorporate chemical-protein interaction data, and thereby allow the association of drugs, toxins and other small molecules with genetic and protein interaction networks. To date, BioGRID has incorporated over 25,000 manually curated small molecule-target interactions from DrugBank (www.drugbank.ca), which encompass more than 2,100 unique human proteins and over 4,300 bioactive molecules. Interestingly, 815 of these human genes are known to be associated with approximately 810 diseases. Visualization of these drug-target associations is facilitated by a new interactive Network Viewer that is embedded in BioGRID search page results. The combination of expertly curated genetic, protein and chemical interaction data into a single resource should facilitate network-based approaches to drug discovery. The entire BioGRID interaction dataset is freely available and may be downloaded in easily computable standardized formats.

11 A Model for Knowledge Extraction from Protein Fingerprint Annotation

Ognyan Kulev

Usually protein databases have been built using resource-consuming manual curation. The influx of high-throughput sequencing data requires a higher level of automated curation solutions for keeping with the ever growing amount. PRINTS, one of the oldest protein databases, is the target of our research, which is aimed at developing of an automated model of inter-database integration and building an enriched fingerprint knowledge base. The main focus of the work is fingerprint annotation, not sequence content.More precise and machine-readable cross references to other databases are made for automatic processing and reasoning; biomedical ontologies, and Gene Ontology in particular, are used in all relevant places. The extracted knowledge is represented using Semantic Web technologies (RDF and OWL) allowing to make complex SPARQL queries that span multiple online databases.The main research objective is to construct an ontology of fingerprints, in which they are described using GO terms, as well as terms from other biomedical ontologies. Current state of this work will be presented and problems will be discussed.

12 The Complex Portal: Annotating complexes made easy

Birgit Meldal, Colin Combe, Josh Heimbach, Henning Hermjakob and Sandra Orchard

The EMBL-EBI Complex Portal (www.ebi.ac.uk/intact/complex), a central service for information on macromolecular complexes, provides manually curated information on stable, protein complexes. We provide unique identifiers, names and synonyms, list of complex members with their unique identifiers (UniProt, ChEBI, RNAcentral), function, binding and stoichiometry annotations, descriptions of their topology, assembly structure, ligands and associated diseases as well as cross-references to the same complex in other databases (e.g. ChEMBL, GO, PDB, Reactome). Complexes are curated into the IntAct database using IntAct tools, rules and quality control and are available for search and download via a dedicated website. They are available for use as annotation objects in a number of resources, including the Gene Ontology where they can be accessed via Protein2GO. We have also developed a novel JavaScript visualisation tool that creates a schematic view of the topology and stoichiometry of each complex. The viewer generates the graphic on the fly, based on the latest database version. Our focus for the next two years lies in increasing content and coverage by importing 'unreviewed' complexes from GO, PDB and Reactome, in addition to increased manual curation, and further improving our graphical options by including 3D structural viewers as well as pathway and expression overlays. A pipeline into InterMine, including incorporation of the complex viewer has been developed, enabling export of organism-specific complexes to model organism resources.This is a collaborative project, which has already been contributed to by groups such as UniProtKB, Saccharomyces Genome Database, the UCL Gene Annotation Team and MINT database. We welcome groups who are willing to contribute their expertise and will make editorial access and training available to you. Individual complexes will also be added to the dataset, on request. Contact us on intact-help@ebi.ac.uk for further information.

13 caNanoLab: Enhancing Retrieval and Sharing of Cancer Nanotechnology Data

Mervi Heiskanen, Stephanie Morris, Sharon Gaheen, Michal Lijowski and Juli Klemm

The US National Cancer Institute (NCI) cancer Nanotechnology Laboratory (caNanoLab) data portal is an online nanomaterial database that allows users to submit and retrieve information on well-characterized nanomaterials used in biomedicine, including composition, in vitro and in vivo experimental characterizations, experimental protocols, and related publications. Currently, more than 1,100 curated nanomaterial records are publicly accessible and can be queried directly from the caNanoLab homepage. The primary customers of the data are the cancer nanotechnology research community including clinicians and the NCI Alliance for Nanotechnology in Cancer. However, the content of caNanoLab is relevant to the broader biomedical research field with interests in the use of nanotechnology for the development of diagnostics and therapeutics. The database structure is based on characterization assays required for clinical regulatory review performed by the Nanotechnology Characterization Laboratory (NCL) and the Centers of Cancer Nanotechnology Excellence (CCNEs) in the NCI Alliance for Nanotechnology in Cancer. The caNanoLab data model was informed by standards such as the NanoParticle Ontology (NPO) and ISA-TAB. Class names and attributes are maintained in the NCI cancer Data Standards Repository (caDSR) and definitions for caNanoLab concepts are maintained in the NCI Thesaurus. The curation of nanotechnology information is accomplished by selecting relevant publications, manually extracting reported text and data, and submitting extracted information into caNanoLab. Curated caNanoLab data are converted to ISA-TAB-Nano files to enable data exchange between individual users or other databases, which is an interest of the caNanoLab team, along with participation in activities focused on the development of standards enabling data exchange and supporting interoperability between databases. The caNanoLab team is also engaged in many activities to better serve the needs of the nanotechnology research community. Activities range from engaging publication vendors to facilitate linkages between publications and nanotechnology databases, to working with other groups to develop data standards and guidelines for data submission and sharing including community-based programs such as the NCI Nanotechnology Working Group (Nano WG) and the National Nanotechnology Initiative (NNI) a federal initiative to develop data standards and deposition guidelines.

14 Creating and Maintaining a Data Archive at the PDB

Jasmine Young, John Westbrook, Stephen Burley, Rcsb Pdb Team and Wwpdb Team

The Protein Data Bank (PDB) is the single global repository for three-dimensional structures of biological macromolecules and their complexes. Over the past decade, the size and complexity of macromolecules and their complexes with small molecules deposited to the PDB have increased significantly. The PDB archive now holds more than 115,000 experimentally determined structures of biological macromolecules, which are all publicly accessible without restriction. These structures, including ribosomes and viruses, provide essential information for understanding biochemical processes and structure-based drug discovery. It is crucial to transform acquired data into readily usable information and knowledge. The PDB archive represents one of the best-curated and most heavily used digital data resources in Biology.Data for each archival entry must be organized and categorized in a meaningful way to support effective data sharing. The PDBx/mmCIF dictionary uses controlled vocabularies to define deposited data items and metadata. Biocuration software tools have been built to use this dictionary to maintain data consistency across the PDB archive. To support scientific advancement and ensure the best data quality and completeness, a working group of community experts in structural biology software works with the wwPDB to enable direct use of PDBx/mmCIF format files across the structure determination pipeline. An overview of creating archive requirements and designing a data model for biological experimental data contributed by multiple data providers will be presented. wwPDB members are RCSB PDB (supported by NSF, NIH, and DOE), PDBe (EMBL-EBI, Wellcome Trust, BBSRC, NIGMS, and EU), PDBj (NBDC-JST) and BMRB (NLM).RCSB PDB, Rutgers, The State University of New Jersey, Piscataway, New Jersey, United States;PDBe, EMBL-European Bioinformatics Institute, Hinxton, United Kingdom;PDBj, Institute for Protein Research, Osaka University, Osaka, Japan;BMRB, BioMagResBank, University of Wisconsin-Madison, Madison, Wisconsin, United States

15 Improvements to the Drosophila anatomy ontology

Marta Costa, David Osumi-Sutherland, Steven Marygold and Nick Brown

The Drosophila anatomy ontology (DAO) defines the broad anatomy of the fruitfly Drosophila melanogaster, a genetic model organism. It contains over 9000 classes, with close to half of these corresponding to neuroanatomical terms. These terms are used by curators when capturing data from papers, and by users when searching for information on FlyBase or Virtual Fly Brain. When the DAO was first developed over 20 years ago, the majority of classes did not include textual information, such as a definition, synonyms or references. These details are essential for curators to be accurate and for users to understand the data. The initial DAO also lacked formalisation, which is critical to ensure correct classification with minimal intervention as the ontology grows. In the last few years, we have made a significant effort to add the missing textual information and to increase the number of inferred classifications. Recently, this work has focused on reviewing the DAO in a systematic manner, making use of the classification into 11 different organ systems, such as muscle, integumentary, adipose, etc. Classes within each organ system have been reviewed together, making it much easier to correct inconsistencies or duplications, and to spot patterns that can be used to write formal definitions.We have so far reviewed classes that belong to 9 of the 11 organ systems in the DAO. This work has increased the number of terms with a definition from 73% to 88%. Future work will focus on completing this effort, by revising the terms in the remaining 2 organ systems.

16 The BioSharing Registry: mapping the landscape of standards and databases resources in the life sciences

Peter McQuilton, Alejandra Gonzalez-Beltran, Allyson Lister, Eamonn Maguire, Philippe Rocca-Serra, Milo Thurston and Susanna-Assunta Sansone

BioSharing (http://www.biosharing.org) is a curated, web-based, searchable portal of three linked registries of content standards, databases, and data policies in the life sciences, broadly encompassing the biological, natural and biomedical sciences. Launched in 2011 and built by the same core team as the successful MIBBI portal, BioSharing harnesses community curation to collate and cross-reference resources across the life sciences from around the world. Every record is designed to be interlinked, providing a detailed description not only on the resource itself, but also on its relations with other life science infrastructures. Serving a variety of stakeholders, BioSharing cultivates a growing community, to which it offers diverse benefits. It is a resource for funding bodies and journal publishers to navigate the metadata landscape of the biological sciences; an educational resource for librarians and information advisors; a publicising platform for standard and database developers/curators; and a research tool for bench and computer scientists to plan their work.With over 1,300 records, BioSharing content can be searched using simple or advanced searches, filtered via a filtering matrix, or grouped via the 'Collection' feature. Examples are the NPG Scientific Data and BioMedCentral Collections, collating and linking the recommended standards and repositories from their Data

Policy for author. Similarly other publishers, projects and organizations are creating Collections by selecting and filtering standards and databases relevant to their work, such as the BD2K bioCADDIE project. As a community effort, BioSharing offers users the ability to 'claim' records, allowing their update. Each claimant also has a user profile that can be linked to their resources, publications and ORCID ID, thus providing visibility for them as an individual. Here, we introduce BioSharing to the International Society of Biocuration, and encourage members to register on the website and claim the record for their database, metadata standard or policy.

17 GenEpiO: The Genomic Epidemiology Application Ontology for the Standardization and Integration of Microbial Genomic, Clinical and Epidemiological Data

Emma Griffiths, Damion Dooley, Melanie Courtot, Josh Adam, Franklin Bristow, Joao A. Carrico, Bhavjinder K. Dhillon, Alex Keddy, Thomas Matthews, Aaron Petkau, Julie Shay, Geoff Winsor, Robert Beiko, Lynn M. Schriml, Eduardo Taboada, Gary Van Domselaar, Morag Graham, Fiona Brinkman and William Hsiao

Genomic Epidemiology, the use of microbial genomic sequences to perform infectious disease outbreak investigation and surveillance, is increasingly being deployed by many public health agencies worldwide. During foodborne outbreaks, contextual information is key for identifying sources of pathogen contamination and exposure. While sequence data usually adheres to a few standardized formats, additional data such as surveillance and exposure information are mostly unstructured and without interoperable standards. Currently, public health workers must rely heavily on computational text and data mining for time-consuming manual curation and analysis of large datasets. A solution providing a framework for integrating these diverse data types is the use of ontologies. Ontologies, well-defined and standardized vocabularies interconnected by logical relationships, support logical reasoning over the data annotated in their terms. Canada's Integrated Rapid Infectious Disease Analysis (IRIDA) project is developing open-source, user-friendly tools for incorporating microbial genomic data into epidemiological analyses to support real-time infectious disease surveillance and investigation. Our research efforts include the development of a Genomic Epidemiology Application Ontology (GenEpiO), which is crucial for epidemiological and genomics data integration.To determine the scope and priorities of GenEpiO development, we interviewed public health stakeholders and domain experts and surveyed reporting forms and databases. User activities, lab management software, information and work flows, exposure tracking and reporting systems were profiled to better characterize users' needs. Community standards were reviewed to determine the utility of different ontologies for fulfilling the identified requirements. Laboratory and epidemiological resources were mined for important fields, terms and descriptors. Our work indicates that no single ontology currently covers all attributes required for a genomic epidemiology program. Furthermore, the very breadth of many ontologies hinders their practical use in real-time by users with little bioinformatics expertise. With this in mind, user profiles and data requirements were harmonized with different ontological standards to create a single resource. An initial OWL file containing metadata fields and terms describing isolate source attribution, clinical data, whole genome sequencing processes, quality metrics, patient demographics/histories/comorbidities and exposures was created adhering to the best practices of the Open Biomedical and Biological Ontology (OBO) Consortium. This application ontology was made more robust through testing in different pathogen surveillance initiatives. Key gaps in domain vocabulary requiring expansion were also identified, e.g. antimicrobial resistance, whole genome sequencing result reporting, food description and epidemiology.IRIDA's GenEpiO is being developed for integrating important laboratory, clinical and epidemiological data fields. Implementation of GenEpiO will facilitate data standardization and integration, validation, interoperability. Improved querying will facilitate automation of many analyses. Since harmonization of the genomic epidemiology ontology can only be achieved by consensus and wide adoption, IRIDA is currently forming an international consortium to build partnerships and solicit domain expertise. The methods developed in this work are also being applied to other datasets such as those associated with the Canadian Healthy

Infant Longitudinal Development (CHILD) study. GenEpiO is a highly anticipated development that will enhance infectious disease investigations, but is also applicable to broader comparative genomic data mining.

18 Sample contextual data integration and management: a marine case study

Petra ten Hoopen, Guy Cochrane and Embric Consortium

An essential part of any useful 'omics' dataset is accurate information on provenance of the material under investigation. Contextual information of a sample, a fundamental unit of a material entity isolated from the surrounding environment and subjected to the investigation, is typically captured as a set of key-value pairs in a sample record, an information artefact about the sampled material. Requirements for the sample contextual information are shaped by the nature of the sample as well as the method of investigation. It is therefore critical that an opinion on the contextual data is formulated in a community of domain experts. Here we describe ongoing efforts of the marine community to harmonise sample contextual data reporting across scientific domains, including the genomic, oceanographic and biodiversity data along with phenotypic traits of aquacultures and characteristics of bioactive natural products originating from marine microbial and microalgae strains promising for blue biotechnology. We will focus on the community-agreed contextual data standardisation and ontologies integration that significantly simplifies reporting of the sample contextual information to public data archives and leads to better discoverability of 'omics' datasets associated with the samples.

19 The Evidence and Conclusion Ontology (ECO): A community resource for representing evidence and supporting assertions

Marcus Chibucos, Suvarana Nadendla, Shoshannah Ball, Dustin Olley, Kimuel Villanova, Dinara Sagitova, Ivan Erill and Michelle Giglio

The Evidence and Conclusion Ontology (ECO) is a community standard for describing biological research evidence in a controlled and structured way. Annotations at the world's most heavily used biological databases (e.g. UniProt, SwissProt, GO, various model organisms, et cetera) are associated with ECO terms, which represent different types of evidence and thus document the supporting evidence for those annotations. Evidence terms described by ECO include experimental and computational methods, author statements curated from the literature, inferences drawn by curators, combinatorial methods, and even statements of provenance. Because ECO is an ontology, where terms with standard definitions are networked to one another using defined relationships, it is possible to conduct selective data queries leveraging the structure of the ontology and automate quality control mechanisms for large-scale data management. A growing number of resources are coming to rely on ECO, and we are actively developing ECO to meet their evidence needs. Here we describe recent developments involving the ECO project and some of its recent collaborations, most notably: (i) release of a new ECO website that contains user documentation, a news section with up-to-date relevant information, visualization tools, and other useful information; (ii) improvements to the ontology structure; (iii) moving ECO development to GitHub; (iv) addition of numerous experimental evidence types; and (v) addition of new evidence classes describing computationally derived evidence, for example "position-specific scoring matrix motif search evidence". At present ECO is used in over 30 applications (of which we are aware). Recently, we have worked with a number of groups to expand representation of evidence in ECO. These groups included SwissProt (diverse experimental assays), UniProt (detection techniques), IntAct (biological system reconstruction), Gene Ontology (logical inference & synapse research techniques), CollecTF (motif prediction), Planteome (genotype-phenotype associations), Ontology of Microbial Phenotypes (microbial assays), and so on. In addition, we have begun collaborating with the Ontology for Biomedical Investigations (OBI) on representing evidence and conclusions, which we hope will ultimately serve as a community model for cross-ontology coordination. As ECO continues to grow as a resource, we are seeking new users and new use cases. Our goal is to become the de facto community standard for representing evidence in biological curation and database annotations. This material (the ontology and related resources) is based upon work supported by the National Science Foundation (of the USA) under Award Number 1458400.

20 Development of a Unified Ontology for Cross-Species Annotation of Genetic Interactions

Christian A. Grove, Rose Oughtred, Raymond Lee, Kara Dolinski, Mike Tyers, Anastasia Baryshnikova and Paul W. Sternberg

Genetic interactions reveal the functional roles genes play in different biological contexts and reflect the buffering capacity of biological systems towards genetic or environmental perturbation. Charting the genetic structure of biological networks is essential for understanding the basis of human health and disease. A network-based approach to the study of disease requires consistent description of genetic interactions in humans and in genetically tractable organisms that serve as instructive models for human biology. Unified descriptors of genetic interactions are also needed to allow accurate comparisons of mutant phenotypes across different species. Toward this end, WormBase (www.wormbase.org) and BioGRID (www.thebiogrid.org) have collaborated on the development of a new Genetic Interaction (GI) Ontology, the goal of which is to unify the nomenclature and interpretation of genetic interactions within the research community and across various Model Organism Databases (MODs). This GI Ontology encapsulates coherent definitions of all known genetic interaction types based on structured descriptors that delineate specific relationships often shared between different interaction types. In order to ensure consistent descriptions across multiple species, the GI Ontology has been developed with support from other major MODs, including SGD, CGD, PomBase, FlyBase, and ZFIN. The GI ontology can be readily combined with species-specific phenotype and tissue ontologies in order to precisely capture the varied effects and contexts of genetic interactions. This compatibility will be extended to the comprehensive cross-species phenotype ontology, UberPheno, as developed by the Monarch Initiative (www.monarchinitiative.org). The BioGRID database will implement the GI Ontology for the curation of genetic interactions in human and model organisms, including yeast, worm, fish and fly. Adoption of standardized GI terms will facilitate the integration of genetic interaction datasets that can now be produced by large-scale CRISPR/Cas9-based screens in human cells and other organisms. Cross-species comparisons of genetic interaction networks will provide key insights into complex human diseases caused by multiple genetic perturbations. The GI Ontology has been integrated as a separate Genetic Interactions branch of the well-established Proteomics Standards Initiative - Molecular Interaction (PSI-MI) ontology (www.obofoundry.org/ontology/mi.html).

21 Developing controlled vocabularies for optimal use in data analysis and visualization

Kirsten Hochholzer, Jana Sponarova and Philip Zimmermann

The advent of genome-wide measurements, such as produced by transcriptomics, has led to the publication of thousands of studies and of the corresponding data files. It is assumed that the availability of these data to the wider research community will facilitate re-analysis and meta-analysis, leading to novel insights that go beyond the primary purpose of these studies. A major challenge in integrating such public studies, however, is the heterogeneity with which they are described. Not only are basic sample information frequently missing, but the descriptions provided often use different vocabularies. Since over ten years, the curation team at the Swiss company NEBION has developed and continuously improved application ontologies describing important biological dimensions such as tissues, genotypes, diseases, cancers, perturbations, or other factors. The main goal of ontology development was to standardize experimental descriptions for optimal use in data analysis. As a result, ontologies were built with minimal redundancy, slim tree depth, and with vocabularies that most biologists understand. In this talk, we will present the main concepts behind our ontologies. In particular, we will discuss how a new Anatomy ontology was developed in collaboration with the CALIPHO group at the SIB in Geneva and successfully applied into an analysis tool like GENEVESTIGATOR.

22
The Future of Curation at dictyBase

Petra Fey, Siddhartha Basu, Robert Dodson and Rex L. Chisholm

The complete dictyBase overhaul and introduction of state of the art software infrastructure will allow curators to begin annotating new biological features and use existing annotations to represent and connect data in novel ways. Curated protein interactions via the Gene Ontology (GO) will be used to represent protein-protein interactions. Curators already privately annotate spatial expression with the Dictyostelium anatomy ontology and we recently started annotating Dictyostelium disease orthologs with their respective disease ontology (DO) terms. The updated database will also allow representing GO annotations with 'GO extensions', which add deeper context to those annotations. In the near future HTML5 technology will revolutionize the way curators add annotations to the database, allowing the direct editing of gene pages. Furthermore, it will open the door for direct community annotations on the gene page for interested users. Basu S, Fey P, Jimenez-Morales D, Dodson RJ, Chisholm RL. dictyBase 2015: Expanding data and annotations in a new software environment. Genesis. 2015 PMID: 26088819

23
UniProtKB: a resource to facilitate enzyme research

Rossana Zaru, Elisabeth Coudert, Kristian Axelsen, Anne Morgat, Uniprot Consortium, Uniprot Consortium and Uniprot Consortium

Enzymes play an essential role in all life processes including metabolism, cell communication and DNA replication. In humans, about 27% of enzymes are associated with diseases making them important targets for drug discovery. In addition, their extensive use in biomedicine and industry highlights the necessity of repositories for enzyme-related data. The UniProt Knowledgebase (UniProtKB) fulfills this need by providing the scientific community with accurate, concise and easy to access information with the aim of facilitating enzyme research.UniProtKB collects and centralises functional information on proteins, with accurate, consistent and rich annotation. For enzymes, which represent between 20-40% of proteomes, UniProtKB provides, in addition to the core annotation, information about EC classification, catalytic activity, cofactors, enzyme regulation, kinetics and pathways all based on the critical assessment of experimental data. Computer-based analysis and, if available, structural data are used to enrich the annotation of the sequence with the identification of active sites and binding sites. Mutagenesis and variants are also annotated. Collectively, they provide valuable information to understand the aetiology of diseases and for the development of medical drugs. By providing accurate annotation of enzymes across a wide range of species, UniProtKB is a valuable resource to make enzyme research easier for scientists and health researchers working in both academia and industry.

24
Expert curation of plant proteins in UniProtKB

Michel Schneider, Damien Lieberherr and Uniprot Consortium

The Universal Protein Resource (UniProt) provides the scientific community with a comprehensive and richly curated resource of protein sequences and functional information. The centerpiece of UniProt is the knowledgebase (UniProtKB) which is composed of the expert curated UniProtKB/Swiss-Prot section and its automatically annotated complement, UniProtKB/TrEMBL.Expert curation combines the manually verified sequence with experimental evidence derived from biochemical and genetic analyses, 3D-structures, mutagenesis experiments, information about protein interactions and post-translational modifications. Besides harvesting, interpreting, standardizing and integrating data from literature and numerous resources, curators are also checking, and often correcting, gene model predictions. For plants, this task is focused on Arabidopsis thaliana and Oryza sativa subsp. japonica.By the end of January 2016, 14'358 manually reviewed entries from Arabidopsis thaliana are present in UniProtKB/Swiss-Prot, including most of the proteins associated with a least one publication containing some functional characterization. Manual expert curation of UniProtKB/Swiss-Prot is complemented by expert-driven automatic annotation of UniProtKB/TrEMBL entries to build a comprehensive, high quality set of proteins covering the complete proteome of Arabidopsis thaliana. This complete set, containing currently 31'477 proteins, is downloadable from the UniProt Web site (http://www.uniprot.org/proteomes/UP000006548). It is based on the latest data provided by the community and we are completing the knowledgebase by importing missing information from EnsemblPlants.We recently started to collaborate with Araport, the Arabidopsis portal, and we provide Araport with all the gene model corrections that we introduced on the bases of our trans-species family annotation. Data from high-throughput proteomics experiments constitute a rich potential source of annotations for UniProtKB. Certified experimental peptides that uniquely match the product of a single gene are used to generate annotations describing post-translational modifications and protein processing events, and to confirm protein existence. Around 2'600 Arabidopsis entries are now containing annotations extracted from 11 reviewed articles describing large scale proteomics experiments.

25 Drosophila curation in UniProtKB

Kate Warner

UniProt provides the scientific community with a comprehensive, high-quality and freely accessible resource of protein sequence and functional information. Each UniProt Knowledgebase (UniProtKB) entry contains as much information as possible and includes core mandatory data (the amino acid sequence, protein name or description, taxonomic data and citation information) as well as widely accepted biological ontologies, classifications, cross-references, and clear indications of the quality of annotation in the form of evidence attribution of experimental and computational data. The fruit fly Drosophila melanogaster has been utilised as a model organism for genetic investigations for over a century. The Drosophila protein annotation program at UniProtKB focuses on the manual annotation of characterised D. melanogaster proteins, and UniProtKB currently contains 3,273 reviewed entries from D. melanogaster. This number continues to increase with each release while existing reviewed entries are revisited and updated as new information becomes available. The UniProt manual curation process for Drosophila will be presented, with emphasis on how UniProtKB entries are structured to aid the retrieval of information by the user.

26 Curation of Zebrafish Models of Human Disease

Yvonne Bradford, Sridhar Ramachandran, Sabrina Toro and Doug Howe

Zebrafish are increasingly used to model and study human disease. Publications are reporting zebrafish mutants, wildtype fish treated with Morpholinos, TALENs, or CRISPRs, or zebrafish that have been exposed to chemical treatments as models of human disease. To facilitate the curation of this information from publications, curation interfaces have been developed at ZFIN (zfin.org) to annotate zebrafish models of human disease utilizing the Disease Ontology (DO, http://disease-ontology.org/) in conjunction with pertinent genotype, sequence targeting reagents and experimental conditions. Each disease model annotation has an evidence code to indicate how it is supported (TAS: Traceable Author Statement; IC: Inferred By Curator), along with a citation to the original publication. Disease model annotations can be viewed on publication pages as well as on disease term pages. Disease term pages include information about the disease term like disease name, synonyms, definition, cross-references and ontological relationships. In addition the disease term page has a section that lists and links to the human genes known to be associated with the disease, OMIM, pages and the corresponding zebrafish orthologs, providing easy access to related information such as zebrafish gene expression data and zebrafish mutant phenotype data. In addition to the data view pages, ZFIN produces download files of these annotations making this information more readily available for the biomedical research community. To enable searching for zebrafish models of human disease, the category "Human Disease" has been added to the ZFIN single box search results, making it easy to find specific disease terms, see relevant genes, and associated models. Likewise, a "Human Disease" filter was added to the single box search Gene results to filter gene sets for those that are associated with a specific human disease. Taken together, the addition of the disease ontology, curated zebrafish models of human diseases, and added data search support will streamline the identification of zebrafish models of human diseases.

27 Rhea, an expert curated resource of biochemical reactions

Anne Morgat, Thierry Lombardot, Kristian B. Axelsen, Lucila Aimo, Anne Niknejad, Nevila Hyka-Nouspikel, Elisabeth Coudert, Steven Rosanoff, Joseph Onwubiko, Nicole Redaschi, Lydie Bougueleret, Ioannis Xenarios and Alan Bridge

Rhea (www.rhea-db.org) is a comprehensive and non-redundant resource of expert curated biochemical reactions described using species from the ChEBI (Chemical Entities of Biological Interest) ontology of small molecules.Rhea has been designed for the functional annotation of enzymes and the description, analysis and reconciliation of genome-scale metabolic networks. All Rhea reactions are balanced at the level of mass and charge. Rhea includes enzyme-catalyzed reactions (covering the IUBMB Enzyme Nomenclature list and additional reactions), transport reactions and spontaneously occurring reactions. Reactions involving complex macromolecules such as proteins, nucleic acids and other polymers that lie outside the scope of ChEBI are also included. Rhea reactions are extensively curated with links to source literature and are mapped to other publicly available metabolic resources such as MetaCyc/EcoCyc, KEGG, Reactome and UniPathway.Rhea reactions are used as a reference for the reconciliation of genome-scale metabolic networks in the MetaNetX resource (www.metanetx.org) and serve as the basis for the computational generation of a library of lipid structures and analytes in SwissLipids (www.swisslipids.org).Here we describe recent developments in Rhea, which include a new website and substantial growth of Rhea through sustained literature curation efforts. At the time of writing, Rhea (release 69, of January 2016) includes 8805 unique approved reactions involving 7688 unique reaction participants, and cites 8500 unique PubMed identifiers.

28 Large-scale inference of gene function through phylogenetic annotation of Gene Ontology terms: case study of the apoptosis and autophagy cellular processes

Marc Feuermann, Pascale Gaudet, Huaiyu Mi, Suzanna Lewis and Paul Thomas

The Gene Ontology (GO) offers a unified framework to describe the roles of gene products in a species-independent manner. Descriptive terms are linked with genes, gene products, or proteins. When direct experimental support is available, genes are associated with the appropriate term and the annotations tagged as having direct evidence. However, the vast majority of proteins will likely never be studied experimentally. To predict the function of these uncharacterized proteins, requires methods that are both efficient and accurate. To this end, the GO consortium has developed PAINT, the Phylogenetic Annotation and INference Tool, based on the protein families generated by the PANTHER protein family classification system. We present here some general observations about phylogenetic annotation inference and primary GO annotations, as illustrated by 36 protein families that participate in two well-characterized cellular processes, apoptosis and autophagy. These processes are well conserved in animals and eukaryotes respectively, and phylogenetic analysis with PAINT reveals their elaboration during evolution. We show that annotation integration via phylogenetic relationships can be used to select high confidence annotations that represent the core functions of protein families. The GO phylogenetic annotation project is extending the coverage of proteins annotated, providing a coherent annotation corpus across a number of representative species. In addition, PAINT improves the quality of the entire set of GO annotations by uncovering discrepancies and inaccuracies in the primary annotations.

29 NavMutPredict: Annotation of the functional impact of mutation in sodium channels.

Aurore Britan, Valerie Hinard, Monique Zahn and Pascale Gaudet

Ion channels allow ions to flow across membranes in all living cells. They play an important role in key physiological processes such as nervous transmission, muscle contraction, learning and memory, secretion, cell proliferation, regulation of blood pressure, fertilization and cell death. In human, 344 genes encode ion channels. Mutations in more than 126 ion channel and ion channel-interacting protein genes have been reported to cause diseases, known as channelopathies. Knowledge on the effect of these mutations is spread throughout the scientific literature. The consolidation of this data on a single platform will help scientists and clinicians to have a source of validated information for diagnosis and therapeutic counseling.The NavMutPredict project focuses specifically on the voltage-gated sodium channel gene family (SCN). The aim is to use all pertinent knowledge concerning mutations in the 9 human sodium channel proteins and their impact on the biophysical properties of the channels. Ultimately, this information should help predicting the pathogenicity of newly discovered genetic variations. To do this, we extract information from the biomedical literature, especially findings related to pathologies and functional data. This data is captured using our biocuration software platform, the BioEditor. This tool allows the capture of structured annotations with a very high level of precision using standardized vocabularies and ontologies, such as GeneOntology, or the Ion Channel ElectroPhysiology Ontology, developed by our group. So far, the BioEditor contains 791 variants found in the SCN proteins and 4127 annotations. All this data will be available to the scientific community via neXtProt and will undoubtedly be a useful resource for a better understanding of ion channel function, essential for understanding channelopathies and developing drugs for new treatments.

30 Exploring autophagy with Gene Ontology

Paul Denny, Marc Feuermann, David Hill, Paola Roncaglia and Ruth Lovering

The Gene Ontology (GO) is a community resource that represents biological knowledge of physiological gene functions through the use of a structured and controlled vocabulary. As part of an ongoing project to improve the representation of specific biological domains in GO, we have focused on autophagy, a process in which cells digest parts of their own cytoplasm and organelles. This allows for both recycling of macromolecules under conditions of cellular stress and remodelling of the intracellular structure during cell differentiation. Well-conserved across species, autophagy is involved in several pathophysiological events relevant to human health, including cancer, metabolic disorders, and cardiovascular and pulmonary diseases; as well as, neurodegenerative processes, such as Parkinson's disease. Autophagy is also implicated in the response to aging and to exercise. We have made significant modifications to the ontology structure and hierarchy for autophagy (www.ebi.ac.uk/QuickGO/GTerm?id=GO:0006914), such as making chaperone-mediated autophagy a direct child of autophagy, rather than a synonym. Some existing terms were renamed to reflect their use in the literature and also new terms were created, e.g. 'protein lipidation involved in autophagosome assembly'. Furthermore, we have created terms such as 'mitophagy in response to mitochondrial depolarization' and 'parkin-mediated mitophagy in response to mitochondrial depolarization', because the recruitment of the Parkin / PINK1 pathway as a result of mitochondrial membrane depolarisation is a key part of the selective autophagy of mitochondria (mitophagy). In some cases, it has been necessary to introduce taxon constraints, which restrict usage of some GO terms to specific taxa; e.g. molecular evidence of classical microautophagy was found only in yeast literature. A similar, yet distinct, process known as late endosomal microautophagy was reported initially in mammals, but it is uncertain whether this should be restricted to multicellular eukaryotes. In addition to improving the ontology, substantial effort was applied to annotate the human and mouse proteins involved in autophagy and the regulation of autophagy. So far we have associated 337 GO terms with 249 human proteins, through the expert curation of 60 papers. It was expected that all of the proteins in the Reactome pathway for macroautophagy (http://www.reactome.org/PathwayBrowser/#R-HSA-1632852) would also be annotated with autophagy-related GO terms; however, we found 8 (out of the 67 proteins in the pathway) discrepancies. These differences were reconciled following further literature searches and GO annotation. Through expansion and refinement of the ontology and the annotation of selected literature, we have substantially enriched and updated the representation of autophagy in the GO resource. This work will support the rapid evaluation of new experimental data, and thus help further elucidate the role of autophagy in disease.Funded by NIH:NHGRI HG22073; Gene Ontology Consortium (PIs: JA Blake, JM Cherry, S Lewis, PW Sternberg and P Thomas) and Parkinson's UK: G-1307.

31 Functional annotation of cardiovascular-related miRNAs using the Gene Ontology

Rachael Huntley, Tony Sawford, Maria Martin, Manuel Mayr and Ruth Lovering

MicroRNA (miRNA) regulation of developmental and cellular processes is a relatively new field of study, however the data generated from such research has so far not been organised optimally to allow inclusion of this data in pathway and network analyses tools. The association of gene products with terms from the Gene Ontology (GO) has proven highly effective for large-scale analysis of functional data, but until recently there has been no substantial effort dedicated to applying GO terms to miRNAs. This lack of functional annotation for miRNAs has been identified as a problem when performing functional analysis, where scientists have to rely on annotation of the miRNA gene targets rather than that of the miRNAs. We have recognised this gap and have started an initiative to curate miRNAs with GO functional annotation, using the Gene Ontology Consortium guidelines for curation of miRNAs http://wiki.geneontology.org/index.php/MicroRNA_GO_annotation_manual. Our plan over the next few years is to build a resource comprising of high-quality, reliable functional annotations for cardiovascular-related miRNAs; annotations that will be a valuable addition to the advancement of miRNA research in this field.Funded by British Heart Foundation grant RG/13/5/30112

32 SABIO-RK database meets user requests

Maja Rey, Ulrike Wittig, Renate Kania, Andreas Weidemann and Wolfgang Muller

SABIO-RK (http://sabio.h-its.org) is a web-accessible, manually curated database that has been established as a resource for biochemical reactions and their kinetic properties with a focus on supporting the computational modelling to create models of biochemical reaction networks. It contains annotations to controlled vocabularies, ontologies and is interlinked with and linked to different databases. A flexible way of exporting database search results in table-like format is provided. Users can tailor their own custom-made export format by selecting properties of the entries in the result set the user wants to export. Both the export and the import of data are possible via SBML format. The already-existing ways of SABIO-RK to collect feedback from users has been extended recently to improve the SABIO-RK database content and to better match user requirements. In case of an empty result, the user is presented with the opportunity to directly request addition of the corresponding data. Similarly, the user can also proactively ask via SABIOs Services for curation of individual papers or e.g. pathway- and organism-associated data. These user requests can be hidden or made visible to the public as curation priority list.SABIO-RK is part of the data management node NBI-SysBio within the de.NBI (German Network of Bioinformatics Infrastructure) program which is a newly-established BMBF-funded initiative having the mission to provide comprehensive first-class bioinformatics services to users in life sciences.

33 The Enzyme Portal - bringing enzyme data together

Joseph Onwubiko, Sangya Pundir, Xavier Watkins, Rosanna Zaru, Steven Rosanoff, Claire O'Donovan and Maria Martin

Enzymes play a vital role in all life processes and are used extensively in biomedicine and biotechnology. Information about enzymes is available in several different Bioinformatics resources, each of which is built with different communities in mind. Researchers are not always aware of how much is available for them and often go to the same resources, missing out on potentially valuable information. The Enzyme Portal brings all of this public information together in one place - it is a unique, invaluable resource for scientists and health researchers working in both academia and industry. EMBL-EBI relaunched the Enzyme Portal in October 2015. Now fully integrated with UniProt and the EBI Search, the Enzyme Portal integrates information from several resources, saving researchers valuable time by providing a concise, cross-linked summary of their enzyme or protein of interest.The Enzyme Portal searches major public resources including the UniProt Knowledgebase (UniProt KB), the Protein Data Bank in Europe (PDBe), Rhea, Reactome, IntEnz, ChEBI and ChEMBL, and provides a summary of catalytic activity, protein function, small-molecule chemistry, biochemical pathways, drug-like compounds, catalytic activity and taxonomy information. It also provides cross-references to the underlying data resources, making it easier to explore the data further. Within the Enzyme Portal, the powerful EBI Search engine can now search a more refined set of enzymes from UniProt, including new enzymes and orthologs in a wide range of species. The sequence search is now based on the enzyme sequence library, using EBI-NCBI Blast. The service offers new search entry points to enzymes according to disease, Enzyme Classification, taxonomy (model organisms) and pathway. There is also a basket functionality, which allows users to store and compare multiple enzymes. The Enzyme Portal's interface is now more intuitive, re-designed based on usability research. Enzyme summaries on the result and entry pages now offer clearer descriptions and set out more complete orthology relationships. Users can easily view enzyme summaries with annotations from PDBe, Rhea, Reactome, IntEnz, ChEBI, ChEMBL and Europe PMC, all of which now provide data to the Enzyme Portal through web services.

34 UniProt Features Viewer: A visual navigation on sequence function annotations

Leyla Jael Garcia, Xavier Watkins, Sangya Pundir, Maria Martin and Uniprot Consortium

Meaningful visualization of sequence function annotation at both gene and protein level is one of the cornerstones in Biology and Bioinformatics. In the protein field, sequence annotations, a.k.a. protein features, are used to describe regions or sites of biological interest; for instance secondary structures, domains, post-translational modifications and binding sites amongst others, play a critical role in the understanding of what the protein does. With the growth in the amount and complexity of biological data, integration and visualization becomes increasingly important in order to expose different data aspects that might be otherwise unclear or difficult to grasp.Here we present the UniProt Features Viewer, a BioJS component bringing together curated and large-scale experimental data. Our viewer displays protein features in different tracks providing an intuitive and compact picture of co-localized elements; initial tracks currently include domains & sites, molecule processing, post translational modifications, sequence information, secondary structure, topology, mutagenesis and natural variants. Each track can be expanded to a detailed view revealing a more in-depth view of the underlying data, e.g., topological domain, trans-membrane and intra-membrane. The variant track offers a novel visualization using a matrix which maps amino acids to their position on the sequence, therefore allowing the display of large number of variants in a restricted space.The UniProt Features Viewer presents tracks under a ruler that represents sequence length for this protein. Anchors located on the left and right sides of the ruler make it easier for users to zoom-in to a particular region. Zooming can also be done via an icon located on top of the categories, by positioning the cursor on top of the features area and scrolling, and by using gestures on desktops and mobile devices. Customization is also possible, particularly, category tracks can be hidden so users can focus on those categories more relevant to their research. Features can be selected by clicking on them; on feature selection additional information such as description, positions and evidence, will be displayed on a tool tip. Some of the type tracks provide a particular tailored view, e.g., for variant data, or use distinctive shapes, e.g., triangles for modified residues or hexagons for glycosylation sites.Modularity and easy integration are core to the UniProt Features Viewer. It has been already integrated into the CTTV website in order to provide a graphical view of proteins, e.g., https://www.targetvalidation.org/target/ENSG00000157764. Other groups such as InterMine (http://intermine.org) have already express their interest in using it. Our viewer has also been tested with users in order to assess usability of the product.We will continue to integrate selected large-scale experimental data, we plan to include proteomics related data, i.e., peptides, as well as antigenic data, i.e., epitope bindings.

35 3D-structure biocuration: a wealth of information in UniProtKB/Swiss-Prot

Ursula Hinz and Uniprot Consortium

Protein 3D-structures provide essential information about protein function in health and disease. UniProtKB/Swiss-Prot biocurators make use of this wealth of data, combining 3D-structure data with information derived from the scientific literature to verify protein function and mode of action, validate enzyme active sites, identify physiologically relevant ligand binding sites and post-translational modifications, and interactions between proteins, or proteins and nucleic acids. This information is shown in a structured format in the UniProtKB/Swiss-Prot entries to facilitate retrieval of specific pieces of information: protein function and subunit structure, cofactor requirements, the role of specific residues, domains and regions, post-translational modifications, membrane topoplogy, etc., with evidence tags to indicate the sources of the information. Information from well-characterized proteins is then propagated to close family members. As a result, out of roughly 550'000 UniProtKB/Swiss-Prot entries, ca. 88'000 contain information about metal-binding sites, 137'000 contain information about the binding sites for nucleotides or other small organic ligands and about 95'000 contain active site annotations, to cite only the most abundant types of annotation.In UniProtKB, cross-references to PDB and PDBSum facilitate access to experimental 3D-structures, while cross-references to Swiss Model repository (SMR) and Protein Model Portal facilitate access to theoretical models.In January 2016, UniProtKB/Swiss-Prot contained 123'700 cross-references to PDB, corresponding to over 22'900 entries, mostly from model organisms. Over 25% (5'700) of the 20'200 human entries have a cross-reference to PDB, and the majority of these have at least one matching literature citation. The situation is similar for other model organisms.

36 miRandola 2016: the latest version of the circulating RNA database.

Francesco Russo, Sebastiano Di Bella, Giovanni Nigita, Federica Vannini, Gabriele Berti, Flavia Scoyni, Alessandro Lagana, Alfredo Pulvirenti, Rosalba Giugno, Marco Pellegrini, Kirstine Belling, Soren Brunak and Alfredo Ferro

Non-coding RNAs (ncRNAs) such as for example microRNAs (miRNAs) are frequently dysregulated in cancer and have shown great potential as tissue-based markers for cancer classification and prognostication. ncRNAs are present in membrane-bound vesicles, such as exosomes, in extracellular human body fluids. Circulating miRNAs are also present in human plasma and serum cofractionate with the Argonaute2 (Ago2) protein and the High-density lipoprotein (HDL). Since miRNAs and the other ncRNAs circulate in the bloodstream in a highly stable, extracellular forms, they may be used as blood-based biomarkers for cancer and other diseases. A knowledge base of non-invasive biomarkers is a fundamental tool for biomedical research.Data is manually collected from ExoCarta, a database of exosomal proteins, RNA and lipids and PubMed. Articles containing information on circulating RNAs are collected by querying PubMed database using keywords such as "microRNA", "miRNA", "extracellular" and "circulating". Data is then manually extracted from the retrieved papers. General information about miRNAs is obtained from miRBase. The aim of miRandola is to collect data concerning RNAs contained not only in exosomes but in all extracellular types functionally enriched with information such as diseases, processes, functions, associated tissues, and their potential roles as biomarkers.Here, we present an updated version of the miRandola database, a comprehensive manually curated collection and classification of extracellular circulating RNAs. The first version of the database has been published in

2012 and it contained 89 papers, 2132 entries and 581 unique mature miRNAs. Now, we have updated the database with 271 papers, 2695 entries, 673 miRNAs and 12 long non-coding RNAs. RNAs are classified into several categories, based on their extracellular form: RNA-Ago2, RNA-exosome, RNA-microvesicles, RNA-HDL and RNA-circulating. Moreover, the database contains several tools, allowing users to infer the potential biological functions of circulating miRNAs, their connections with phenotypes and the drug effects on cellular and extracellular miRNAs.miRandola is the first online resource which gathers all the available data on circulating RNAs in a unique environment. It represents a usufeul reference tool for anyone investigating the role of extracellular RNAs as biomarkers as well as their physiological function and their involvement in pathologies. miRandola is constantly updated (usually once a year) and the online submission system is a crucial feature which helps ensuring that the system is always up-to-date.The future direction of the database is to be a resource for all the potential non-invasive biomarkers such as cell-free DNA, circular RNA and circulating tumor cells (CTCs). miRandola is available online at: http://atlas.dmi.unict.it/mirandola/References1) Francesco Russo, Sebastiano Di Bella, Giovanni Nigita, Valentina Macca, Alessandro Lagana, Rosalba Giugno, Alfredo Pulvirenti, Alfredo Ferro. miRandola: Extracellular Circulating microRNAs Database. PLoS ONE 7(10): e47786. doi:10.1371/journal.pone.00477862) Francesco Russo*, Sebastiano Di Bella*, Vincenzo Bonnici, Alessandro Lagana, Giuseppe Rainaldi, Marco Pellegrini, Alfredo Pulvirenti, Rosalba Giugno, Alfredo Ferro. A knowledge base for the discovery of function, diagnostic potential and drug effects on cellular and extracellular miRNAs. BMC Genomics 2014, 15(Suppl 3):S4. doi:10.1186/1471-2164-15-S3-S4

37 5S ribosomal RNA database

Maciej Szymanski, Andrzej Zielezinski and Wojciech Karlowski

Ribosomal 5S RNA (5S rRNA) is a conserved component of the large subunit of all cytoplasmic and the majority of organellar ribosomes in all organisms. Due to its small size, abundance and conservation 5S rRNA was used for many years as a model molecule in studies on RNA structure, RNA-protein interactions as well as a molecular marker for phylogenetic analyses. 5SRNAdb is the first database that provides a high quality reference set of ribosomal 5S RNAs (5S rRNA) across three domains of life.To reduce the redundancy of the data set each individual database record represents a unique 5S rRNA sequence identified for particular species. Identical sequences from the same species and deposited under distinct accession numbers in the GenBank/ENA databases are collapsed into single records with links to to the original GenBank records. All of the records in the database are available in the form of manually curated structural sequence alignments in which each column corresponds to a particular position in the general model of the secondary structure of 5S rRNA. Each individual sequence record or a consensus sequence of multiple records is visualized as a secondary structure diagram showing the most general model based on all sequences from a particular group or from the set of records defined by the user. To make the comparison of alignments and general structure models possible, both the alignments and secondary structure diagrams are produced on the templates including all positions present in the master alignments of all sequences from respective taxonomic domains (i.e. Archaea, Bacteria and Eukaryota and organelles). The content of the alignment can be customized by users. The sequences can be added to the alignment by providing record identifiers or by performing database search. The nucleotide statistics and secondary structure models are dynamically recalculated to match the current set of sequences in the updated alignment. Alignments can also be generated from from scratch by adding subsequent search results. The user interface of the 5S rRNA database was designed to incorporate several solutions enhancing the efficiency of the data mining. All browse and search results are shown in separate collapsible windows allowing users to adjust the amount of information visible on each page. The database is available on-line at http://combio.pl/5srnadb/

38 Curation of RNA-Seq differential expression analysis with functional annotation to determine

cancer-specific glycogene expression profiles

Hayley Dingerdissen, Radoslav Goldman and Raja Mazumder

Although it has been well-documented that glycosylation regulates the development and progression of cancer through involvement in fundamental processes like cell signaling, invasion, and tumor angiogenesis, much needs to be developed for a full understanding of the cancer glycome and glycoproteome. Glycosylation of proteins can be altered during malignant progression with respect to the glycan structures but also in their associations with the glycoproteins (sites of attachment and their occupancy). Glycosylation is one of the most prominent post-translational modifications, predicted to affect more than half of all human proteins, but understanding of its full extent is incomplete, especially in the cancer-context. Furthermore, because glycosylation is a coordinated enzymatic pathway, observation of altered glycosylation products could be rationalized in terms of changes in enzyme expression during neoplastic transformation. To study the interplay of proteins involved in glycosylation, both glycoproteins and glycosyltransferases, we conducted genome-wide next-generation sequencing (NGS) analysis of RNA-Seq samples labeled as liver hepatocellular carcinoma (LIHC) from The Cancer Genome Atlas (TCGA). Using BioXpress, a curated gene expression and disease association database, we identified all genes for the given type of cancer which are differentially expressed between corresponding tumor and normal pairs. We then cross-referenced this gene list with a comprehensive list of glycan binding proteins (GBPs) and glycosyltransferases. To generate the glycosylation-specific gene list, we first retrieved the curated list of human GBPs and glycosyltransferases from the Consortium for Functional Glycomics (CFG) Functional Glycomics Gateway (http://www.functionalglycomics.org/fg/). We then retrieved all human UniProtKB/SwissProt entries with keywords Glycoprotein or Glycosyltransferase. Reported differentially expressed genes were then filtered to report those genes involved in glycosylation. Additionally, we retrieved expression information from BGEE, the database for Gene Expression Evolution, and we further reduced the genes of interest to those designated to be orthologous across a subset of organisms to study the cancer-associated glycogenes from an evolutionary perspective. To demonstrate the critical function of curation in studies of this type, we re-analyzed the subset of genes with predicted glycosylation sites derived from the NetNGlyc server (http://www.cbs.dtu.dk/). From this simple comparison, we can readily see that the completeness, and perhaps more crucially the correctness, of glycosylation-related annotations directly impacts our ability to derive functional understanding from such an analysis. We plan to apply this pipeline to a comprehensive pan-cancer study to determine possible glyco-profiles associated with gene expression in different types of cancer, and to automate the entire pipeline through the BioXpress engine.

39 To be folded, to be unfolded or to be aggregated?

Oxana Galzitskaya

In my talk I will describe three possible states of the protein molecules and the corresponding databases and servers for predictions of the disordered and amyloidogenic regions, folding nucleus and handedness. Disordered regions play important roles in protein adaptation to challenging environmental conditions. Flexible and disordered residues have the highest propensities to alter the protein packing. Therefore, identification of disordered/flexible regions is important for structural and functional analysis of proteins. We created the first library of disordered regions based on the known protein structures from the clustered protein data bank. Recently we analyzed the occurrence of the disordered patterns in 122 eukaryotic and bacterial proteomes to create the HRaP database. Amyloid fibrils formation in organs and tissues causes serious human diseases. Therefore identification of protein regions responsible for amyloid formation is one of important tasks of theoretical and experimental investigations. Recently the role of a mirror image conformation as a subtle effect in protein folding has been considered. The understanding of chirality both in protein structures and amyloid suprastructures is an important issue in molecular biology now. We are the first who have investigated the relationship of the protein handedness with the rate of protein folding.

40 **Beyond GWAS - connecting phenotypes, genotypes, and biological knowledge**

Clay Birkett, David Matthews, Peter Bradbury and Jean-Luc Jannink

The Triticeae Toolbox (T3) triticeaeatoolbox.org is a database for wheat, barley, and oat that contains genotype and phenotype data used by plant breeders. To allow breeders to select markers for developing new germplasm we have done meta-analysis on all the trials. We preformed Genome-wide association studies (GWAS) on each of the 334 phenotype experiments, 55 genotype trials, and 147 traits. The genotypes where imputed using Beagle version 4 using a 1.3 million SNP haplotype map for better resolution. The resulting quantitative trait loci (QTL) are identified by location in the reference genome and in JBrowse genome browser. The QTLs are prioritized by the gene annotation. The tables provide links to EnsemblPlant for identification of protein and comparative genomics. The website will also be using QTLNetMinner ondex.rothamsted.ac.uk/QTLNetMiner to integrate gene information with annotation, biochemical pathway, gene expression, comparative genomic, and publications. The QTLNetMinner also provides us with a network viewer to visualize the connections of the integrated information.

41 **The BioGRID Interaction Database: Curation strategies and new developments**

Lorrie Boucher, Rose Oughtred, Jennifer Rust, Christie Chang, Bobby-Joe Breitkreutz, Nadine Kolas, Lara O'Donnell, Chris Stark, Andrew Chatr-Aryamontri, Kara Dolinski and Mike Tyers

The Biological General Repository for Interaction Datasets (BioGRID) (http://www.thebiogrid.org) is an open source database for protein and genetic interactions, protein post-translational modifications and drug/chemical interactions, all manually curated from the primary biomedical literature. As of February 2016, BioGRID contains over 1,052,000 interactions captured from high throughput data sets and low throughput studies experimentally documented in more than 45,800 publications. Comprehensive curation of the literature has been completed for protein and genetic interactions in the budding yeast S. cerevisiae and the fission yeast S. pombe and protein interactions in the model plant A. thaliana. The comprehensive curation of human interaction data is currently not feasible due to the vast number of relevant biomedical publications each month. To address this in part, we have taken the approach of themed curation of interactions implicated in central cell biological processes, in particular those implicated in human disease. In order to enrich for publications that contain relevant interaction data, we are using state-of-the-art text mining methods, which have effectively doubled the rate and coverage of our manual curation throughput over the past five years. To date, we have curated themed human interaction data in the ubiquitin-proteasome system (UPS), the autophagy system, the chromatin modification (CM) network and the DNA damage response (DDR) network. With the recent development CRISPR/Cas9 genome editing technology, the stage is set to explosively expand the landscape of human genetic interaction data through genome-wide phenotypic and genetic interaction screens. In conjunction with WormBase and other model organism database partners, we have developed a genetic interaction ontology to allow rigorous annotation of genetic interactions across all species, including humans. We are currently building a dedicated portal with BioGRID to allow interrogation of high-throughput human genetic interaction data. A curation pipeline has also been established to capture chemical/drug interaction data from the literature and other existing resources (see poster by Oughtred et al.). All BioGRID data is archived as monthly releases and freely available to all users via the website search pages and customizable dataset downloads. These new developments in data curation, along with the intuitive query tools in BioGRID that allow facile data mining and visualization should help enable fundamental and applied discovery by the biomedical research community.

42 **Curation of chip-seq datasets for collection of transcription factor binding motifs**

Ilya Vorontsov, Ivan Kulakovskiy, Ivan Yevshin, Anastasiia Soboleva, Artem Kasianov, Haitham Ashoor, Wail Ba-Alawi, Vladimir Bajic, Yulia Medvedeva, Fedor Kolpakov and Vsevolod Makeev

Knowledge on transcription factors (TF) and their binding sites (TFBS) provide basis for a wide spectrum of studies in regulatory genomics, from reconstruction of regulatory networks to functional annotation of transcripts and sequence variants. While TFs may recognize different sequence patterns in different conditions, it is pragmatic to have a single generic model for each particular TF as a baseline for practical applications. We provide the HOmo sapiens COmprehensive MOdel COllection (HOCOMOCO, http://hocomoco.autosome.ru) containing DNA binding patterns for nearly 6 hundreds of human and 4 hundreds of mouse TFs. ChIP-Seq data appears to be the most informative data source on TF binding in vivo. Yet, a ChIP-seq peak does not warrant the tested protein to bind DNA directly, without any mediators, and provides only an approximate location of the protein binding site. In vitro technologies like HT-SELEX warrant direct binding but usually yield less accurate and biased motif specificity. The precise location of binding sites in genomic segments can be predicted by computational methods based on binding motif models such as positional weight matrices (PWMs) or dinucleotide PWMs (diPWMs).

43 Caenorhabditis Annotation in the UniProtKB Database

Hema Bye-A-Jee

The UniProt Knowledgebase (UniProtKB) endeavours to provide the scientific community with a comprehensive and freely accessible resource of protein sequence and functional information on a large variety of model organisms to further scientific research. The nematode worm, Caenorhabditis elegans (C.elegans), is a transparent roundworm, which is approximately 1mm in length and has a relatively short life cycle. The use of C.elegans as a versatile model organism for studying gene and protein function in complex biological processes is adopted by thousands of scientists worldwide. It was the first multicellular organism to be sequenced and was also the organism in which RNA interference was first discovered. Such scientific breakthroughs have paved the way for numerous other large scale sequencing and knockout projects in other multicellular organisms. Nevertheless, C. elegans still remains an essential model organism and practical genetic tool in which to study gene and protein function. The UniProtKB Caenorhabditis protein annotation project focuses on the manual annotation of experimentally characterised C. elegans proteins and also contains entries from other Caenorhabditis species including briggsae, drosophilae, remanei and vulgaris. To date, there are 3,671 manually curated protein entries for C. elegans with this number continuously increasing through on-going annotation. Each comprehensive UniProtKB entry presents primary protein data (including amino acid sequence, gene and protein names and taxonomic data), amalgamated data from a range of sources (including scientific literature, model organism databases and sequence analysis tools) as well as biological ontologies, classifications and cross-references in order to provide an accurate overview. In particular, UniProtKB works closely with both the nematode worm research community and with WormBase, the database of the biology and genome of C. elegans and related nematode species, to ensure that UniProtKB presents detailed and current proteomes, sequences and functional annotation of nematode proteins in a clear, informative and organised manner in order to facilitate and promote further nematode research.

44 Scientific literature and media annotation using the iCLiKVAL browser extension

Naveen Kumar and Todd Taylor

The wide availability of web-based desktop and mobile applications has tremendously accelerated the pace of online content generation and consumption. Researchers are producing and consuming more and more online content than ever, in various forms such as documents, images, audios, videos, software codes, and many more. There is countless information hidden in multimedia contents, though it is often not discoverable due to lack of relevant structured and curated annotations.iCLiKVAL (Interactive Crowdsourced Literature Key-Value Annotation Library) is a web-based application (http://iclikval.riken.jp) that uses the power of crowdsourcing to collect annotation information for all scientific literature and media found online. The iCLiKVAL browser extension is an open source easy-to-use tool, which uses the iCLiKVAL API to save free-form, but structured, annotations as "key-value" pairs with an optional "relationship" between them. The basic idea is to map the online media to a unique URI (Uniform Resource Identifier) and then to assign semantic value to the media to make information easier to find and allow for much richer data searches and integration with other data sources.The browser extension facilitates users to bookmark the content or to mark for "review later". It can be used in offline mode, and the data is automatically synchronized when it is back online. To use this browser extension, users need to be registered with the iCLiKVAL web application. Currently, it is available for the Google Chrome browser, and later it will be available for other popular cross-platform browsers.

45 The Impact of Literature Curation and Annotation on Citation Rates

Tanya Berardini, Leonore Reiser, Michael Lauruhn and Ron Daniel Jr.

Literature curation by model organism databases (MODs) results in the interconnection of papers, genes, gene functions, and other experimentally supported biological information, and aims to make research data more discoverable and accessible to the research community. That said, there is a paucity of quantitative data about how literature curation affects access and reuse of the curated data. One potential measure of data reuse is the citation rate of the article used in curation. If articles and their corresponding data are easier to find, then we might expect that curated articles would exhibit different citation profiles when compared to articles that are not curated. That is, what are the effects of having scholarly articles curated by MODs on their citation rates? To address this question we have been comparing the citation behavior of different groups of articles and asking the following questions: (1) given a collection of 'similar' articles about Arabidopsis, is there a difference in the citation numbers between articles that have been curated in TAIR and ones that have not, (2) for articles annotated in TAIR, is there a difference in the citation behavior before vs. after curation and, (3) is there a difference in citation behavior between Arabidopsis articles added to TAIR's database and those that are not in TAIR? Our data indicate that curated articles do have a different citation profile than non-curated articles that appears to result from increased visibility in TAIR. We believe data of this type can be used to quantify the impact of literature curation on data reuse and may also be useful for MODs and funders seeking incentives for community literature curation. This project is a research partnership between TAIR (The Arabidopsis Information Resource) and Elsevier Labs.

46 Improved discoverability and connectivity of all scientific media through iCLiKVAL annotations

Todd Taylor and Naveen Kumar

Scientific media comes in a variety of languages and formats, including journal articles, books, images, videos, blog and database entries, and so on. In the case of textual media, there is often additional information, such as tables, figures and supplementary data, associated with or embedded in the text. While there are many good resources for browsing, searching and annotating some of this media, there is no single place to search them all at once, and generalized search engines such as Google do not allow for the type of comprehensive and precise searches that researchers require. One could argue that any scientific media that is on the web is therefore connected, but much of it remains offline (e.g., books) or is inaccessible (not open source, only found in libraries, etc.) and is therefore neither discoverable nor connected. To address these issues, we created iCLiKVAL (http://iclikval.riken.jp/), an easy-to-use web-based tool that uses the power of crowdsourcing to accumulate annotation information for all scientific media found online (and potentially offline). Annotations in the form of key-relationship-value tuples (in any language), added by users through a variety of options, make information easier to find and allow for much richer data searches by basically linking all media together through common terminology.Users can create or join common interest groups, both public and private, to annotate related media together as a community. Users can also create and edit their own controlled vocabulary lists, or import established vocabularies such as Medical Subject Headings (MeSH) and Gene Ontology (GO) terms, and then easily select which lists they would like to use for auto-suggest terms in the key, value and relationship fields. Within the user groups, vocabulary and bookmark lists can be shared so that everyone uses the same standards and works towards a common goal. In addition, we have implemented a notification center, several customization options, and a one-stop annotations feature where users can view and edit all of their own annotations. Most of the pages used for tracking progress, such as annotations, bookmarks, search history, reviews and vocabularies, are searchable, sortable and filterable, so users can quickly find what they are looking for.Our goal is the ability to add key-value pairs to any type of scientific media. While iCLiKVAL was initially limited to the annotation of PubMed articles, we recently added the capability to curate media from YouTube, Flickr and SoundCloud. And, to really broaden our scope, anything with a digital object identifier (DOI) can now be annotated using iCLiKVAL, allowing for the inclusion of hundreds of millions of media objects and more. While the interface is very intuitive and easy to use on almost every browser and platform, we have also created a Chrome Browser extension that allows any non-password protected online media to bookmarked and annotated, to facilitate the linking of all scientific media. The iCLiKVAL database is completely searchable, and all of the collected data is freely available to registered users via our API.

47 Defining a repertoire of innate immunity genes contributing to intracellular defense against pathogens

Weihua Chen, Antonio Rausell, Fellay Jacques, Amalio Telenti and Evgeny Zdobnov

Intracellular defense against pathogens is a fundamental component of human innate immunity. However, the numbers of genes defined as being part of innate immunity are largely inconsistent among existing annotation datasets. Therefore, there is a need for better criteria to identify the subset of genes acting as intracellular effectors. In this study, we aim to approach this using machine learning methods. Our primary analysis with a shallow implementation of a classifier of innate immunity genes suggested that better cross-validation results can be obtained with the use of innate immunity genes that are covered by more existing annotation datasets as the true-positives, highlighting the importance of high-quality training data. We thus have launched a crowd-sourcing curation project in order to define two high-quality training datasets for the machine learning, one consists of innate immunity genes, the other consist of none-immunity-related genes. In total 2000 genes were randomly selected, among which about half are putative innate immunity genes covered by public datasets, the more number of datasets a gene is covered, the higher likelihood it will be included in the datasets; while the other half are putative none-immunity-related genes that were randomly selected from the other human genes excluding those that are covered by any existing innate-immunity databases, those that are homologous to these genes and those that are adaptive-immunity-related. For each gene we collected their annotations from public databases, expression changes upon interferon stimuli, Gene Ontology classification and gene knockout phenotypes. Based on these information, curators will be asked to vote if it is definitely innate-immunity-related, unsure or definitely not. Each curator will be asked to vote randomly on 250 genes out of the 2000; each gene will be voted by at least 11 curators. Then we will calculate a consensus for each curation task based the "majority" rule and assign genes accordingly to the "positive" and "negative" datasets. We will train the machine learning algorithms on the two datasets with various genetic, genomic and evolutionary features that we have already collected for all human genes. Through feature selection we should be able to obtain a list of features that are informative in distinguishing innate-immunity genes from others. In addition, we will apply the resulting model to other genes to search for putative new innate immunity genes, and submit them for further experimental validation.

48 PomBase Literature Curation

Antonia Lock, Midori Harris, Kim Rutherford, Valerie Wood and Jurg Bahler

PomBase obtains its highest-quality data by manual curation of the fission yeast literature, which provides experimentally supported annotations representing gene structure, function, expression and more. Approximately 5000 papers suitable for manual curation have been published on fission yeast to date, of which about 2100 have been fully curated.To supplement the work of its small staff of professional curators, PomBase has developed a community curation model that enables researchers to participate directly in curating data from their own publications. As of April 2015, the fission yeast community has contributed annotations for over 260 publications. Community curation improves the visibility of recent publications, and enables researchers and professional curators to work together to ensure that PomBase presents comprehensive, up-to-date and accurate representation of published results.Furthermore, because PomBase is one of only three databases that provide manual literature curation for fungal species, electronic data transfer of high-confidence fission yeast annotations to other fungal species is an essential source of functional data for the latter. Community contributions to PomBase therefore support research not only within the fission yeast community, but also throughout the broader community studying fungi.

49 Curating Multi-Allele Phenotypes in PomBase

Midori Harris, Antonia Lock, Kim Rutherford, Mark Mcdowall and Valerie Wood

PomBase, the model organism database for fission yeast, makes the comprehensive and detailed representation of phenotypes one of its key features. We have made considerable progress in developing a modular ontology, the Fission Yeast Phenotype Ontology (FYPO), for phenotype descriptions, and in making phenotype annotations for single mutants available. Canto, the PomBase community curation tool, provides an intuitive interface for curators and community users alike to associate alleles with FYPO terms and supporting metadata such as evidence, experimental conditions, and annotation extensions that capture expressivity and penetrance. The PomBase web site displays phenotype annotations on gene pages, and supports FYPO term searching by name or ID. We are now extending the PomBase phenotype annotation resources to annotate phenotypes observed in double mutants, triple mutants, etc. The Chado database underlying PomBase supports annotation of specific alleles, singly or in combinations, by associating phenotypes with genotypes which in turn link to their constituent alleles. Extensive additions and adaptations of the Canto phenotype annotation interface enable curators and researchers to capture multi-allele phenotype data and metadata. We invite comments on extending the PomBase gene page display and search options to accommodate and use the new data.

50 GenomeRNAi: A Phenotype Database for Large-scale RNAi Screens

Esther E. Schmidt, Oliver Pelz, Benedikt Rauscher and Michael Boutros

RNA interference (RNAi) represents a powerful strategy for the systematic abrogation of gene expression. High-throughput screening experiments, performed for a wide variety of underlying biological questions, result in the description of loss-of-function phenotypes across many fields in biology. These phenotypes represent a major contribution to the annotation of gene function.The GenomeRNAi database holds information on RNAi phenotypes and reagents, aiming to provide a platform for data mining and screen comparisons. The database is populated by manual data curation from the literature, or via direct submission by data producers. Structured annotation guidelines are applied for curatorial review. Where possible, a controlled vocabulary is defined for given data fields. At present, the database contains 452 experiments in human, and 201 experiments in Drosophila, providing more than 2,5 million individual gene-phenotype associations in total. A recent major contribution has been the Broad Institute's Achilles project with genome-wide shRNA screening data for 216 different cancer cell lines. GenomeRNAi also integrates information on efficiency and specificity for more than 400,000 RNAi reagents, obtained by running a quality assessment pipeline on a regular basis.The GenomeRNAi website (www.genomernai.org) features search functions (by gene, reagent, phenotype or screen details), as well as options for browsing and downloading experiments. Further features are a "frequent hitter" page, and a functionality for the overlay of genes sharing the same phenotype, onto known gene networks provided by the String database (www.string-db.org). GenomeRNAi data are well integrated with external resources, providing e.g. mutual links with FlyBase, GeneCards and UniProt. GenomeRNAi functional data have also been incorporated into the FlyMine tool. Given the sharp increase in data volume we are currently working on visualisation options for a more intuitive, overview-like display of the data. These will be presented at the conference.

51 International Mouse Phenotyping Consortium: 5 years of annotated mouse functional data

Luis Santos, Ann-Marie Mallon and Mouse Phenotyping Informatics Infrastructure

The International Mouse Phenotyping Consortium (IMPC) was founded with the purpose of building a truly comprehensive, functional catalogue of the mouse genome by means of high-throughput phenotyping and classification of single-gene knock-out mouse lines, allowing discovery of new mouse models for human diseases. As of January 2016, over 95000 specimens belonging to 3274 KO lines (and controls) had been phenotyped, and around 1,260,000 experimental procedures were performed. To effectively and usefully benefit from this wealth of data, new robust statistical analysis methodologies have been developed and implemented to test for the existence of phenodeviants, and to automatically classify them using the Mammalian Phenotype ontology developed by Mouse Genome Informatics (MGI) at Jackson Laboratory.From the onset of the project until the present day, the strategies and mechanisms used in annotating the data have changed and evolved; a mix of manual and automated annotation is used, with the former being performed at the very specialised level and only in certain data contexts - such as anatomical or image observation. To date, we have identified 1461 phenotypes, all annotated according to the IMPC phenotyping pipeline that includes as part of its standardized protocols the ontological outcome(s) to be assigned when a statistically significant outlier strain is detected for a phenotype test . This process, developed with biologists from all the IMPC partner centres, is an important step to ensure that the automatic annotation of phenodeviancy reflects the underlying biological system that is altered.Apart from contributing to maintaining and expanding MGI's MP terms catalogue, the IMPC also employs programmatic solutions to infer relationships between mouse and human phenotype representations, making the most use of resources such as MP, OMIM, Orphanet and PhenoDigm to discover new mouse models of disease and provide insights into disease mechanism.

52 Essential requirements for community annotation tools.

Monica C Munoz-Torres, Chris Mungall, Nathan Dunn, Seth Carbon, Heiko Dietze, Nicole Washington, Jeremy Nguyen Xuan, Paul Thomas and Suzanna Lewis

Scientific research is inherently a collaborative task; in our case it is a dialog among different researchers to reach a shared understanding of the underlying biology. To facilitate this dialog we have developed two web-based annotation tools: Apollo (http://genomearchitect.org/), a genomic feature editor, designed to support structural annotation of gene models, and Noctua (http://noctua.berkeleybop.org/), a biological-process model builder designed for describing the functional roles of gene products. Here we wish to outline an inventory of essential requirements that, in our experience, enable an annotation tool to meet the needs of both professional biocurators as well as other members of the research community. Below are the general requirements, beyond specific functional requirements, that any annotation tool must satisfy. These include: - Real time updates to allow geographically dispersed curators to conduct joint efforts; - Immediate communication between curators through parallel chat mechanisms; - Rigorously documenting the experimental or inferential basis for all of the annotations that are made with credit assigned through citations; - Well supported history mechanisms providing the ability to comment on versions, browse versions to see different edits and commentary, and revert to earlier versions; - Providing different levels of permissions for users and administrators, for example so that a curator might "doodle" within their own work area before releasing their version for feedback from others;- Offering incentives for adoption, such as facilitating the publication process;- Prompt responsiveness to users' requests and informative documentation, and dedicated resources for training and user support, from online seminars to video tutorials to repositories with teaching materials; - Functional stability and ease of migrating forward when new software is released; - And, most importantly, a publishing mechanism, such that biocurators and other contributors receive credit for their insights and contributions to our collective understanding of biology.

53 Annotating GigaDB; have your say.

Christopher Hunter, Xiao Sizhe, Peter Li, Laurie Goodman and Scott Edmunds

The current incarnation of GigaDB relies upon the paid curators at GigaScience to read, understand and annotate each entry manually. This gives the best possible annotations, but obviously cannot be sustained with ever increasing numbers of submissions. We are investing significant time and effort to enable more people to provide those annotations, such as authors (through a submission wizard) who obviously know their data best of all; the reviewers whom have an active interest in the articles and are therefore in a prime position to suggest relevant annotations; as well as the general users - either casually clicking through or seriously making use of the data. Everyone can add something, if only they were given the tools to do so.Here we present the current and already active addition of hypothes.is, a web curation layer, to the GigaDB pages which allows anyone to add comments and annotations to our pages, either publicly or to make private notes about the data for themselves or their "group". In addition we will show some future initiatives that we are planning to help keep data held in GigaDB up to date with the as much metadata as is available, making the data as discoverable and as useful to as many people as possible.

54 Sustainable biocuration of functional annotation at the European Nucleotide Archive

Richard Gibson, Blaise Alako, Clara Amid, Ana Cerdeno-Tarraga, Iain Cleland, Neil Goodgame, Petra ten Hoopen, Suran Jayathilaka, Simon Kay, Rasko Leinonen, Xin Liu, Swapna Pallreddy, Nima Pakseresht, Jeena Rajan, Marc Rossello, Nicole Silvester, Dmitriy Smirnov, Ana Luisa Toribio, Daniel Vaughan, Vadim Zalunin and Guy Cochrane

The European Nucleotide Archive (ENA; http://www.ebi.ac.uk/ena) hosts, maintains and presents nucleotide sequence data along with associated sample and experimental metadata. Functional annotation has always been hand-in-hand with the storage of traditional sequence records, providing an interpretation of genetic features together with the sequence itself. Such information enables discoverability of the data, regardless of whether the annotation is supported by experiment or inference alone. As nucleotide sequencing has become cheaper and more productive, particularly in the area of whole genome sequencing, the number of assembled sequences, and therefore functional annotations, has continued to grow at an unprecedented rate. ENA has been addressing these challenges with developments over the last six years. These include: (1) an increase in the number of checklists useful for simple and common functional annotations (such as bacterial genes, rRNA genes and phylogenetic markers); (2) provision of editable skeleton files (known as 'templates'), useful for the more advanced user in submitting more complex annotation; (3) support for the submission of viral genomes through the Assembly Pipeline; and (4) extension of automatic biological rule-based validations at the time of submission. Such ongoing changes are not only providing a smoother experience and faster turnaround for the user but are re-shaping the role of the ENA biocurator to a more sustainable biocuration.

55 Community curation efforts at FlyBase

Gary Grumbling, Jim Thurmond, Josh Goodman, Thom Kaufman and Flybase Consortium

FlyBase has had a successful community curation tool in place for several yearscalled Fast-Track Your Paper (FTYP). It is a web application that presentsusers with a series of forms and creates curation records from their input.These records then feed into our curation pipeline with no special processingnecessary. The records serve to help triage publications for further curationby FlyBase staff and to expedite the linking of genes to publications in ourdatabase. This web application works in conjunction with another system thatemails the authors of recently published Drosophila papers requesting that theyfill out the forms. We have consistently gotten an approximately fifty percentresponse rate to these emails over about five years time leading to some seventhousand curation records created by users. We will present more usagestatistics and experiences from the deployment of the tool over this time. Thetechnical details of our rewrite of the original tool released this past yearwill also be presented. We will also discuss possible future directions of ourcommunity curation efforts at FlyBase.

56 National Bioscience Database Center: the activity for integration and usage promotion of life science databases in Japan

Shigeru Yatsuzuka, Jun-Ichi Onami, Tomoe Nobusada, Atsuko Miyazaki, Hideki Hatanaka and Toshihisa Takagi

National Bioscience Database Center (NBDC; http://biosciencedbc.jp/en/) was founded in 2011, as a core institution for the integration of life science databases in Japan. We took over our mission from Integrated Database Project (2006-2010). For integration and usage promotion of life science databases scattered among many research institutes, we have conducted the following activities: - Strategic planning for database research and development - Enhancement of life science databases - Sharing research data - International cooperationAnd we have released, expanded and sophisticated the following services:Integbio Database Catalog (http://integbio.jp/dbcatalog/?lang=en):The catalog includes basic information of greater part of life science databases in Japan and major life science databases around the world. It consists of URL, database maintenance site, category, organism, operational status, and so on. We curate those information according to our consistent curation policy. Now we prepare to release the RDF formatted version of the catalog.Life Science Database Cross Search (http://biosciencedbc.jp/dbsearch/index.php?lang=en):Life Science Database Cross Search is a searching service across more than 160 databases which include literatures and patent publications. This system is composed of a lot of servers in 5 organizations of 4 ministries. To realize distributed search among remote organization, we adopted full text search engine, Hyper estraier. Now, we started changing this engine to ""Elasticsearch"" so as to resolve growing index data. Curators play an important role in adding a new database to this system. Curators investigate the crawling region in each database, predict the URIs of each entry, and classify attribute information. So, cross search curators are needed to have two backgrounds, information technology and life science. We constructed a reasonable workflow for advanced curation of cross search, and we achieved to add new database tenfold faster than before. Now, user can also search ""Deep-web"" data (even Google cannot find) from this cross search system. This system can be used as an infrastructure of comprehensive database search.Life Science Database Archive (http://dbarchive.biosciencedbc.jp/index-e.html):The archive collects life science databases scattered among many research institutes. We will stably keep and maintain them over the long term. To help users to understand databases, we provide detailed metadata of databases and curate them. Each metadata links to the information of researchers (ORCID, researchmap), articles (PubMed, J-GLOBAL) and funds (Life science projects in Japan, J-GLOBAL). To promote reuse, databases in the archive are published under Creative Commons Attribution-Share Alike (CC BY-SA) in principle.Integrated Search:Connecting databases organically enables users to search them with complicated conditions. To realize it: - We are RDFizing Integbio Database Catalog and databases in Life Science Database Archive. - We have released RDF portal (http://integbio.jp/rdf/). It collects life science data in RDF format. - We are developing tools for RDF search in collaboration with Database Center for Life Science (DBCLS; http://dbcls.rois.ac.jp/en/).

57 IMGT(R) biocuration of IG and TR in IMGT/LIGM-DB and IMGT/GENE-DB

Geraldine Folch, Joumana Jabado-Michaloud, Marine Peralta, Melanie Arrivet, Imene Chentli, Melissa Cambon, Pascal Bento, Souphatta Sasorith, Typhaine Paysan-Lafosse, Patrice Duroux, Veronique Giudicelli, Sofia Kossida and Marie-Paule Lefranc

IMGT(R), the international ImMunoGeneTics information system(R), http://www.imgt.org, is the global reference in immunogenetics and immunoinformatics [1]. By managing the extreme diversity and complexity of the antigen receptors of the adaptive immune response, the immunoglobulins (IG) or antibodies and the T cell receptors (TR) [2, 3] (2.1012 different specificities per individual), IMGT(R) is at the origin of immunoinformatics, a science at the interface between immunogenetics and bioinformatics [4]. IMGT(R) is based on the concepts of IMGT-ONTOLOGY [5] and these concepts are used for expert annotation and standardized knowledge in IMGT/LIGM-DB, the IMGT(R) database of IG and TR nucleotide sequences from human and other vertebrate species and in IMGT/GENE-DB, the IMGT(R) gene and allele database. The IMGT/LIGM-DB biocuration pipeline of IG and TR sequences includes IMGT/LIGMotif, for the analysis of large genomic DNA sequences, and IMGT/Automat, for the automatic annotation of rearranged cDNA sequences. Analysis results are checked for consistency, both manually and by using IMGT(R) tools (IMGT/NtiToVald, IMGT/V-QUEST, IMGT/BLAST, etc.). The annotated sequences are integrated in IMGT/LIGM-DB and include the sequence identification (IMGT keywords), the gene and allele classification (IMGT nomenclature), the constitutive and specific motif description (IMGT labels in capital letters, no plural), the translation of the coding regions (IMGT unique numbering) [4, 5]. For genomic IMGT/LIGM-DB sequences containing either an IG or TR variable (V), diversity (D) or joining (J) gene in germline configuration or a constant (C) gene, the gene and allele information is entered in IMGT/GENE-DB. In parallel, the IMGT Repertoire is updated (Locus representations, Gene tables and Protein displays (for new genes), Alignments of alleles (for new and/or confirmatory alleles)) and the IMGT reference directory [1, 4] is completed (sequences used for gene and allele comparison and assignment in IMGT(R) tools (IMGT/V-QUEST, IMGT/HighV-QUEST for next generation sequencing (NGS), IMGT/DomainGapAlign) and databases (IMGT/2Dstructure-DB, IMGT/3Dstructure-DB). An IMGT/GENE-DB entry also provides information on the rearranged cDNA and gDNA entries (with links to IMGT/LIGM-DB) and on the three-dimensional structures (with links to IMGT/3Dstructure-DB). IMGT/GENE-DB is the official repository of IG and TR genes and alleles. IMGT(R) gene names were approved by HGNC and endorsed by WHO-IUIS, the World Health Organization (WHO)-International Union of Immunological Societies (IUIS) Nomenclature Subcommittee for IG and TR. Reciprocal links exist between IMGT/GENE-DB and HGNC, NCBI and Vega. The definition of antibodies published by the WHO International Nonproprietary Name (INN) Programme is based on the IMGT(R) concepts [6], and allows easy retrieval via IMGT/mAb-DB query [1, 4]. The IMGT(R) standardized annotation has allowed to bridge the gaps for IG or antibodies and TR between fundamental and medical research, veterinary research, repertoire analysis, biotechnology related to antibody engineering, diagnostics and therapeutical approaches.[1] Lefranc M-P et al. Nucleic Acids Res 43:413-422 (2015) PMID: 25378316, [2] Lefranc M-P, Lefranc G. The Immunoglobulin FactsBook (2001), [3] Lefranc M-P, Lefranc G. The T cell receptor FactsBook (2001), [4] Lefranc M-P. Front Immunol 5:22 (2014) PMID: 24600447, [5] Giudicelli V, Lefranc, M-P. Front Genet 3:79 (2012) PMID: 22654892, [6] Lefranc M-P. mAbs 3(1):1-2 (2011) PMID: 21099347.

58 Gene Groups in FlyBase

Helen Attrill and Giulia Antonazzo

The FlyBase 'Gene Groups' resource provides sets of Drosophila melanogaster genes that share common features. These groups are currently restricted to well-defined, easily delimited groupings such as evolutionary-related gene families (e.g. actins, Wnts), subunits of macromolecular complexes (e.g. ribosome, SAGA complex), sets of genes whose products share a common molecular function (e.g. phosphatases, GPCRs, ubiquitin ligases) and gene complexes (e.g. Enhancer of split complex). Gene Groups are manually curated from the primary literature ensuring that the groups are of a high-quality and fully attributed. FlyBase has integrated the building of this resource with a review of gene annotation data. First, for each group the membership is checked to ensure that the group is complete and represents the current research literature and genome annotation. Second, the Gene Ontology (GO) annotation is reviewed to ensure that groups of genes are annotated with terms that reflect their core common biology. Third, a review of the gene nomenclature is conducted to improve consistency and reflect community usage. Gene Group data in FlyBase are presented in the form of Gene Group Reports that include convenient download and analysis options, together with links to equivalent gene groups at other databases. This new resource will enable researchers with diverse backgrounds and interests to easily view and analyse acknowledged sets of fly genes.

59 Taxonomy, a can of worms

Sandrine Pilbout, Teresa Manuela Batista Neto and Nicole Redaschi

Taxonomy encompasses the description, identification, nomenclature and classification of organisms. Unfortunately the scientific literature and data repositories are plagued by incorrect taxonomic assignments, with organism names that are erroneously assigned, ambiguous, out-dated, or simply misspelled, errors that complicate data integration and exploitation. It is therefore crucial to build and maintain taxonomy databases that provide standardized nomenclature and identifiers, and to promote their usage in the research and bioinformatics community. There are many taxonomy standardization efforts that all rely on expert curation. At UniProt we employ the NCBI taxonomy database as our base for taxonomic descriptions. We systematically review and curate the taxonomic assignment for every organism which enters UniProtKB/Swiss-Prot and discuss and resolve inconsistencies and errors with the NCBI taxonomists on a daily basis. This informal collaboration is a significant contribution to maintaining an important resource that is used by many other bioinformatics databases.

60 Evidence attribution in UniProtKB: linking protein information to its source

Michele Magrane, Cecilia Arighi, Sylvain Poux, Nicole Redaschi, Maria Martin, Claire O'Donovan and Uniprot Consortium

UniProt provides the scientific community with a comprehensive, high-quality and freely accessible resource of protein sequence and functional information. It facilitates scientific discovery by organising biological knowledge and enabling researchers to rapidly comprehend complex areas of biology. Information in the UniProt Knowledgebase (UniProtKB) is integrated from a range of sources such as scientific literature, protein sequence analysis tools, other databases and automatic annotation systems to provide an integrated overview of available protein knowledge. As the data are derived from multiple disparate sources, it is important that users can easily trace the origin of all information. To achieve this, UniProt makes use of a subset of evidence codes from the Evidence Ontology to indicate data origin. This system allows users to trace the source of all information and to differentiate easily between experimental and computationally-derived data. An overview of the evidence codes used, how are these are displayed to users and how they can be used to retrieve specific categories of data will be presented. All data are freely available from www.uniprot.org.

61 Annotation and evidence based functional curation of RefSeq prokaryotic genomes

Kim Pruitt, Stacy Ciufo, Michael Dicuccio, Daniel Haft, Wenjun Li and Kathleen O'Neill

The National Center for Biotechnology Information (NCBI) has developed a robust prokaryotic genome annotation pipeline which is offered as a service to GenBank submitters and is used to annotate RefSeq prokaryotic genomes. NCBI's annotation pipeline integrates annotation quality standards which were developed in part through a series of microbial assembly and annotation workshops held by NCBI. These workshops defined annotation quality criteria for complete genomes, standards for reporting support evidence, and protein naming protocols which are used by UniProt, GenBank, and RefSeq for eukaryotic and prokaryotic genomes. In 2015 the RefSeq project implemented a number of changes impacting the prokaryotic genomes data set. These include: a) development of a new framework for evidence based protein naming, name curation, and evidence tracking; b) increased collaboration to improve protein names; c) completion of the transition to a new data model for managing prokaryotic protein sequences; d) expanded assembly and annotation quality testing; e) increased capacity in NCBI's prokaryotic genome annotation pipeline; and f) comprehensive reannotation of all RefSeq prokaryotic genomes. These developments expand on the established annotation QA, evidence, and name standards. As a result, RefSeq provides significant annotation consistency which facilitates comparative genomics. Our more rigorous QA criteria resulted in the suppression of over two thousand RefSeq prokaryotic genomes that do not meet these criteria, thus ensuring continued high quality of the RefSeq prokaryotic genomes dataset. Based on our new infrastructure to support protein name curation, we have established evidence-based and tracked names for approximately 25% of the RefSeq proteins thus far. Curation of protein names and name evidence uses a multi-faceted approach that leverages Hidden Markov Models (HMMs), domain architectures, available curated name data from Swiss-Prot and the Enzyme Commission, collaboration, and curation by NCBI staff. We use HMMs from several sources including TIGRfams and Pfams, and are creating new HMMs (NCBIfams) when further refinement is needed. The RefSeq group collaborates with NCBI's Conserved Domains Database (CDD) group to provide functional names based on reviewed domain architectures. We collaborate with individual scientists and expert databases to provide the best names for some classes of proteins. NCBI staff curate name information both at the level of the support evidence (HMMs), and at the level of individual RefSeq protein names. Database tables support tracking the name update history, the biocurator or collaborator source of the update, and all support evidences for the name at the time of the update. We plan to start reporting functional evidence information on RefSeq protein records in 2016. The presentation will summarize the current RefSeq prokaryotic genomes process flows for genome annotation and curation/collaboration, our current assembly and annotation quality criteria, examples of annotation improvements resulting from collaboration and NCBI staff curation, and a proposal for reporting sets of functional evidence in the context of NCBI sequence displays.

62 Data curation methods at the ENCODE DCC

Jason Hilton, Cricket Sloan, Ben Hitz, Esther Chan, Jean Davidson, Idan Gabdank, J Seth Strattan and J Michael Cherry

The Encyclopedia of DNA elements (ENCODE) project, currently in its 10th year of production scale, is a collaborative effort toward cataloging genomic annotations. The research institutes within the consortium have produced data from more than 5,000 experiments using a variety of techniques to study the structure, regulation, and transcription profiles of human and mouse genomes. Furthermore, the ENCODE site (https://www.encodeproject.org/) has incorporated data generated through other projects involving genomic assays of fly and worm genomes. All of the data displayed on the ENCODE portal first passes through the Data Coordination Center (DCC) for basic validation and metadata standardization. As the amount of data that goes through the DCC continues to grow exponentially, it is necessary to increase the attention and effort given to the curation and organization of metadata. At the ENCODE DCC, we have made vast use of a variety of tools to aid in capturing and integrating experimental details. The ENCODE DCC's active contribution to ontology databases and our use of ontologies as a means to standardize metadata allows for the ease of identifying and comparing related experiments. Additionally, the ENCODE project is structured such that the DCC interacts with production centers from the proposal of the project all the way through data submission. This results in constant and efficient metadata modeling, as well as high quality data even before publication. Here, we discuss the strategies employed by the ENCODE DCC to maximize accessibility to epigenomic data and analysis.

63 NCBI Metadata Resources for Data Discovery

Karen Clark, Tanya Barrett and Ilene Mizrachi

The National Center for Biotechnology Information (NCBI) hosts two resources BioProject and BioSample, which facilitate the capture of structured metadata for diverse biological research projects and samples represented in NCBI's archival databases. BioProject (http://www.ncbi.nlm.nih.gov/bioproject/) is an organizational framework to access information about a research initiative. Links to data submitted to NCBI and other International Nucleotide Sequence Database Collaboration (INSDC) archives are aggregated in BioProject for easy access to the submitted data. Publications, grant information and descriptive information are also captured. BioProjects can exist in a hierarchical structure to represent studies that are part of a larger research initiative. The BioSample (http://www.ncbi.nlm.nih.gov/biosample/) database stores descriptive information or metadata about the biological materials used in studies for which data is submitted to NCBI and other INSDC archives. BioSample packages represent broad categories of sample types and help guide submitters to provide the appropriate descriptive attributes. Rich sample descriptions are important for data users to fully understand the context of the experiments and allow them to more fully interpret the outcomes. As with BioProject, links to submitted data are presented in BioSample for access to all of the data generated from a particular sample. BioProjects and BioSamples are created as part of the data submission process or linkages can be asserted to previously registered records. For users, both of these resources provide an entry point to the submitted data stored in the archives and allow users to access data based on queries based on specific attributes of interest to their research.

64 Analysis of classification algorithms applied to some prokaryotic organisms

Milana Grbic

Classification algorithms are intensively used in discovering new information in large datasets. In this work several classification algorithms are applied to a dataset of prokaryotic organisms. A comparative analysis of the algorithms is performed based on the variations of the data types, dataset dimensions and presence/absence of the attributes. The analysis indicates which of the considered classification models is most suitable for this dataset. Results obtained in this analysis can be useful in further researches devoted to grouping of the considered organisms.

65 SourceData: Making Data Discoverable

Robin Liechti, Lou Goetz, Sara El-Gebali, Nancy George, Isaac Crespo, Ioannis Xenarios and Thomas Lemberger

With the huge body of information already published and the rapid increase in the number of papers indexed in PubMed every year, it is becoming increasingly important to be able to search in a systematic way through biological data. Compounding this issue is the fact that most data are published in the form of figures. In figure format, the experimental data, which provides the evidence for scientific claims, are not machine-readable and, therefore, are neither re-usable nor discoverable. In this way valuable data might be lost or unnecessarily repeated.To address these issues, we have initiated the SourceData project (http://sourcedata.embo.org) with the aim to provide a scalable, structured way to annotate and search through data resulting from hypothesis-driven research in cell and molecular biology.To this end, we have developed a biocuration interface for data editors, to extract and integrate machine-readable metadata, which can be added to figures and their underlying source data. This process can be integrated into the publication workflow and hence does not require authors to alter their submission process. Taking advantage of the information provided by authors in figure legends, data editors identify biological entities in experiments and specify their role in the experimental design. Furthermore these entities are disambiguated, by linking them to entries in existing biological databases. Once a paper is curated by data editors, a secondary validation interface is presented to authors so that they can review the process and ensure the result is an accurate reflection of their data.The resulting semantic information is used to build a searchable 'scientific knowledge graph' that is objectively grounded on published data and not on the potentially subjective interpretation of the results. The resulting SourceData search platform and SmartFigure viewer enable targeted searches of scientific papers based on their data content, thus complementing current keyword-based search strategies. By enhancing the discoverability of research data, SourceData aims at stimulating new discoveries by helping scientists to find, compare and combine data over multiple studies.

66 Using structured output learning in automated protein function annotation

Jovana Kovacevic and Predrag Radivojac

The task of structured output learning is to learn a function that enables prediction of complex objects such as sequences, trees or graphs for a given input. One of the problems where such methods can be applied is protein function prediction. With growing number of newly discovered proteins and slow and expensive experimental methods for their functional annotation, the necessity for fast and accurate tools for protein function prediction has risen in the past several years. Reliable information on protein function is especially important in context of human diseases, since many of them can occur due to alteration of function upon mutation. In protein function prediction, the aim is to find one or more functions that it performs in a cell according to its characteristics such as its primary sequence, phylogenetic information, protein-protein interactions, etc. The space of all known protein functions is defined by a directed acyclic graph known as Gene Ontology (GO), where each node represents one function and each edge encodes a relationship such as is-a, part-of, etc. Each output, on the other hand, represents the subgraph of GO, consistent in a sense that it contains a protein's functions propagated to the root of the ontology.In this research, we developed structured output predictor that determines protein function according to the histogram of 4-grams that appear in the protein's sequence. The predictor is based on the machine learning method of structural support vector machines (SSVM), which represents generalization of the well-known SVM optimizers on structured outputs. Adjusting SSVM to this specific problem required the development of an optimization algorithm that maximizes an objective function over the vast set of all possible consistent subgraphs of protein functional terms as well as careful choice of loss functions.To investigate the influence of the organism that the protein originates from on quality of the protein function prediction, we constructed 5 prediction models trained on proteins of single organisms (human, rat, mouse, E.coli and A. thaliana) and cross-tested each model on proteins from each other organism. The results obtained are comparable with the last CAFA (Critical Assessment of Function Annotation) competition results - for rat, mouse and A. thaliana the results are in the top 15%. As expected, best results for an organism are obtained by the model trained on proteins of the organism itself, except for mouse and rat for which human proteins-trained model performed better. The results suggested dependence of the developed predictor on the volume and quality of training data, and confirmed protein function similarity of evolutionarily close organisms.

67 neXtA5: Accelerating Annotation of Articles via Automated Approaches in neXtProt.

Luc Mottin, Julien Gobeill, Emilie Pasche, Pierre-Andre Michel, Isabelle Cusin, Pascale Gaudet and Patrick Ruch

Introduction: Biological databases serve an important need of storing, organizing and presenting research data to the research community, as the sheer amount of data published make it essentially impossible for an expert to have the corpus knowledge of pertinent to her/his research field. The rapid increase in the number of published articles also poses a challenge for curated databases to remain up-to-date. To help the scientific community and database curators to deal with this issue, we have developed an application - neXtA5 - which prioritizes the literature for specific curation requirements. Methods: Our system, neXtA5, is composed of two main elements. The first is a module based on text-mining that weekly annotates MEDLINE over 5 axes: diseases, the three aspects of the Gene Ontology (GO), and protein interactions. Afterwards, it stores findings in our local database, BioMed. Additional entities such as the species or the chemical compounds are also extracted and displayed to further facilitate the work of curators. The second element is exploiting an Information Retrieval component, which uses the density of entities in abstracts, to prioritize the publications. The ranking function is performed independently on the five different annotation axes.Results: Based on the Text REtrieval Conference evaluation model, we increased precision from 0.28 to 0.42 for the disease axis, and from 0.36, to 0.44 for the GO Biological Process axis. We are currently working on optimizing parameters to improve Molecular Function, Cellular Components and the protein-protein interaction axis.Conclusion: Our application aims to improve the efficiency of annotators by providing a semi-automated curation workflow. A user-friendly interface powered with a set of JSON web services are currently being implemented into the neXtProt annotation pipeline. Available on: http://babar.unige.ch:8082/neXtA5

68 Benchmarks for measurement of duplicate detection methods for bioinformatics databases

Qingyu Chen, Justin Zobel and Karin Verspoor

Duplication is a key data quality problem in many domains. It is related to redundancy (near-identical instances), incompleteness (fragmentary records), and inconsistency (records with contradictory information) - issues that undermine the efficiency of data retrieval and the accuracy of answers to queries. It is problematic in bioinformatics databases, where data volumes are rapidly increasing. The high data volume makes fully manual approaches, recognized as the most precise way to determine whether a pair is duplicate, infeasible. Automatic approaches may be feasible, but methods to date have examined only limited types of duplicates under simple assumptions. A further fundamental issue is that the definition of 'duplicate' is context dependent. A pair considered as duplicates by one community, or for one task, may not necessarily be so in another context; a duplicate detection method that achieves high performance in a dataset gathered under restrictive assumptions may not necessarily be effective on another dataset.We have collected records that can be regarded as duplicates under a range of assumptions, and created related benchmarks. We built three DNA sequence database benchmarks, based on information drawn from a range of resources, including information derived by mapping between databases. Each benchmark has distinct characteristics. We quantitatively measure these characteristics and argue for their complementary value in evaluation of duplication detection techniques. The benchmarks collectively contain a vast number of validated biological duplicates; the largest has nearly half a billion duplicate pairs. They are also the first benchmarks targeting the primary nucleotide databases. The records cover the 21 most studied organisms in molecular biology research. Our quantitative analysis shows that duplicates in the different benchmarks, and in different organisms, have different characteristics. It is thus unreliable to evaluate duplicate detection methods against any single benchmark. For example, the benchmark derived from Swiss-Prot mappings identifies more diverse types of duplicates, showing the importance of expert curation, but is limited to coding sequences. Overall, these benchmarks form a resource that we believe will be of tremendous value for development and evaluation of duplicate detection methods. They also represent the diversity and complexity of duplication in biological databases.

69 The BEL Information Extraction Workflow (BELIEF): Updates and Evaluation

Sam Ansari, Justyna Szostak, Sumit Madan, Sven Hodapp, Philipp Senger, Marja Talikka, Juliane Fluck and Julia Hoeng

Ever-increasing scientific literature enhances our understanding on how toxicants impact biological systems. In order to utilize this information in the growing field of systems biology and toxicology, the published knowledge must be transformed into a structured format highly efficient for modelling, reasoning, and ultimately high throughput data analysis and interpretation. Consequently, there is an increasing demand from systems biologists and toxicologists to access such knowledge in a computable format, here biological network models.The Biological Expression Language

(BEL) is a machine- and human-readable language that represents molecular relationships and events as semantic triples: subject-relationship-object. These triples are called BEL statements. BEL statements associated with their evidence are encapsulated into the large BEL document. The BEL document also captures additional information such as article references, PMID of the scientific article and a large annotation dataset that accurately defines the context of knowledge such as the organism, tissue and disease state. BEL statements can computationally be assembled to biological network models. To facilitate encoding and curation of biological findings in BEL, a BEL Information Extraction workFlow (BELIEF) was developed. BELIEF contains a text mining pipeline for the automatic generation of BEL compliant knowledge statements and a web-based curation interface - the BELIEF Dashboard - that facilitates manual curation of the automatically generated BEL statements. The text mining pipeline is UIMA-based and accommodates several named entity recognition processes and relationship extraction methods in order to detect concepts and BEL relationships from any text resource. The BELIEF Dashboard reuses the output of the BELIEF pipeline to create a web-based interface and facilitate the manual curation task. Although BEL itself enables curation given the human-readability of the syntax, BELIEF simplifies the curation process by highlighting named entities and disambiguating gene and protein names. In addition to the new features of BELIEF, we also present the competitive performance results based on the BioCreative V BEL track evaluation. The BELIEF pipeline automatically extracts normalized concepts with the best F-score of 76.8%. The detection of full relationships and entirely correct statements was achieved with the F-score of 43.1% (second-best) and 30.8% (highest). The participation at the Interactive task (IAT) track in BioCreative V revealed a Systems Usability Scale (SUS) of 67. Given the complexity of the task for untrained users this score certifies a high usability for BELIEF. In conclusion, BELIEF simplifies the curation process and facilitates the construction of biological network models that can be fully contextualized and used for the interpretation of systems biology and systems toxicology data. This workflow is currently being further developed to be used in an industrial setup for product risk assessment.

70 Using machine learning to evaluate the functional annotation consistency of five reference species

Julien Gobeill, Luc Mottin, Emilie Pasche and Patrick Ruch

We used machine learning in order to put to the test the functional annotation consistency of the five most curated species in the Gene Ontology Annotation (GOA) database: homo sapiens, saccharomyces cerevisiae S288c, rattus norvegicus, drosophila melanogaster and mus musculus. The studied task was to automatically infer a list of Gene Ontology (GO) concepts from some MEDLINE citations. Machine learning systems aim to exploit a knowledge base containing already annotated contents (gene products, GO concepts, PMIDs), in order to propose a ranked list of GO concepts for any not yet curated input PMID. In other words, such systems learn from the knowledge base in order to reproduce the curators' behavior. In this study, we used the GOCat system, our local GO classifier that implements a k-Nearest Neighbors algorithm. For the design of the knowledge base, we used GOA in order to collect the MEDLINE citations associated with a gene and a set of GO concepts; we thus populated the knowledge base with an equal amount of 9,000 MEDLINE abstracts for each species, along with their associated GO concepts. 1,000 supplementary abstracts (200 for each species) were used to build the test set: their associated GO concepts were hidden, and GOCat had to recover them. As the knowledge base and the test set come from the same source, the ability to assign a GO concept should directly depend on the consistency of the annotations. GOCat outputs a list of generated GO concepts ranked by confidence scores. The performances of GOCat are measured using two standard metrics. Top Precision (P0) is the fraction of the proposed GO concepts that are correct in the top of the GOCat ranking. Recall at 20 (R20) is the fraction of the GO concepts that were in GOA and that were successfully propoed by GOCat in the first 20 ranks. The baseline results for the whole test set (1,000 abstract) are 0.45 for P0 and 0.57 for R20. Yet, significant differences are observed across species. Homo sapiens and saccharomyces cerevisiae S288c abstracts obtain better results than the baseline:

respectively 0.52 for P0 (+16%) and 0.60 for R20 (+5%), 0.45 for P0 (equal) and 0.67 for R20 (+18%). In contrast, rattus norvegicus and mus musculus obtain results lower than the baseline: respectively 0.38 for P0 (-16%) and 0.52 for R20 (-9%), 0.42 for P0 (-7%) and 0.47 for R20 (-18%). Drosophila melanogaster performances are similar to the baseline. These results show that the homo sapiens functional annotation in GOA is more consistent than the rattus norvegicus and mus musculus ones.

71 Text2LOD: building high-quality linked open annotation data concerning biological interests

Yasunori Yamamoto, Shinobu Okamoto, Shuichi Kawashima, Toshiaki Katayama, Yuka Nakahira-Yanaka, Hiroko Maita and Sumiko Yamamoto

Text2LOD: building high-quality linked open annotation data concerning biological interestsBiological features and metadata of organisms are mainly described in literature, and therefore it is difficult for computational methods to analyze and understand a plethora of genomic data in terms of biological aspects. Many efforts have been taken to extract biological knowledge from literature by using text mining technologies and by developing domain specific dictionaries and ontologies. However, knowledge of some biologically interesting aspects haven't been fully extracted and stored in structured formats such as environments or place where each organism grows and lives. In this situation, we are developing a system to automatically extract them from full texts of papers that describe the genome sequence of an organism.Currently, we are building gold standard data sets focusing on several biologically interesting aspects, that is, habitat environments, sampling places, cell sizes, growth temperature and pH of (targeted) organisms/microbes/microbial species. Three domain experts are annotating papers that were obtained from PMC Open Access subsets. The total number of annotated papers is 2627 as of writing, and that of annotations is 3718. The most annotated aspect is living environments, and the numbers are 1395 and 1517, respectively.While we continue to annotate papers, we are also developing the extraction system. We employ a supervised machine learning approach and template based extraction methods depending on aspects. Our goal is to provide such datasets as Linked Open Data that can be accessed easily from both human and computers without registrations. Database Center for Life Science (DBCLS) provides a platform where you can easily search and browse multiple biological aspects of organisms called TogoGenome. The dataset of each aspect uses Resource Description Framework and can be accessed by SPARQL in addition to the TogoGenome site. Therefore, we plan to provide our datasets through TogoGenome.We discuss what we've learnt and future works.

72 Overview of the BioCreative V Track 4 Challenge: Extraction of Causal Relationships in Biological Expression Language

Juliane Fluck, Sumit Madan, Tilia Ellendorff, Theo Mevissen, Simon Clematide, Adrian van der Lek and Fabio Rinaldi

With the ever-increasing number of published scientific literature, the necessity of automatic extraction of biological knowledge to build biological networks has become indispensable. In order to approach the task of automatic extraction and network generation, the improvement of already available methods as well as the development of new methods and systems is crucial. Track 4 at BioCreative V offered a challenge to develop, adapt, and evaluate text mining systems using material of manually curated biological networks, represented in Biological Expression Language (BEL). BEL is a standard knowledge representation language for systems biology that allows to express molecular and cellular relationships in form of nanopublications, also named BEL statements. The language has been specifically designed to be both human-readable and machine-computable. BioCreative V track 4 included two specific and independent tasks, evaluating two complementary aspects of the problem. Task 1 evaluated text mining systems that are capable of automated BEL statement construction from a given text snippet. For task 2, the systems needed to suggest 10 additional text snippets for a given BEL statement. A BEL statement consists of multiple biological terms, functions, and relations. For task 1, our chosen evaluation methodology considered these fragments of information expressed by BEL statements and evaluated the systems at each of these structural levels. The aim of this evaluation strategy was to help identify the key features and characteristics of each system. For the evaluation of task 2, the text snippets for a given BEL statement were manually assessed by experts on three different levels of increasing strictness.To perform the tasks, participating systems needed to be capable of high-quality recognition of biological terms, their normalization to database entries, the extraction of the relationships between terms, and the transformation of all this information into BEL syntax. The systems used state-of-the-art methods based on dictionary-lookup, rules derived from expert knowledge and advanced machine learning to perform well in the task. At term level, the best systems scored around 69%. For relation extraction, up to 72.7% F-score was reached. In contrast, the results on extracting protein function terms were relatively poor, around 30% F-score. False or missing function assignments were also one of the main reasons for the low score (18.2%) of full BEL statement extraction. The performance increased significantly to 25.6% when gold standard entities were provided by the organizers. For task 2, F-scores between 39.2% and 61.5% were reached depending on the strictness of the applied criterion. In summary, track 4 at BioCreative V showed that manually curated BEL networks can be used as training data to develop new text mining methods and systems. The training and test data as well as the evaluation environment is available for further development of these systems, and future extensions of the annotated data are planned. The resulting systems can hopefully be deployed to assist biocuration for network generation in the area of systems biology.

73 BioCreative V: a community-wide effort for the evaluation of text mining and its relevance for biomedical curation

Cecilia Arighi, Kevin Cohen, Donald C. Comeau, Rezarta Islamaj Dogan, Juliane Fluck, Lynette Hirschman, Sun Kim, Martin Kralliner, Zhiyong Lu, Fabio Rinaldi, Alfonso Valencia, Thomas Wiegers, W. John Wilbur and Cathy Wu

BioCreative is a community-wide effort for evaluating text mining systems applied to the biological domain. BioCreative has been running since 2004, providing the premier evaluation forum for this domain. Previous editions of BioCreative have included tasks such as recognition of gene mentions and their normalization to database identifiers, identification of protein-protein interactions, function annotation from text using GO terms, and extraction of chemicals, drugs and diseases and relations among them. BioCreative has spearheaded several innovations in the field, including the development of corpora, evaluation methodologies and interoperability standards (BioC).The tasks in BioCreative V, in addition to addressing suggestions from the biocuration community, included several novel tasks. The following specific tasks were evaluated:- Track 1 (BioC track- Collaborative Biocurator Assistant Track): Interoperability of components assembled for a curation task. Developers were invited to provide complementary text mining modules that could be seamlessly integrated into a system capable of assisting BioGRID curators. The simple BioC format ensured interoperability of the different components.- Track 2 (CHEMDNER-patents track) Processing of chemical entities in patent data, a resource type currently underrepresented in public annotation databases.- Track 3 (CDR track) Extraction of chemical-disease relations from the literature, using the Comparative Toxicogenomics database (CTD) as a potential curation target.- Track 4 (BEL track) Extraction of fragments of pathway networks, in particular causal networks, in a formal language known as Biological Expression Language (BEL), and the extraction of evidence sentences from the literature for given BEL statements.- Track 5 (IAT track) The curator-centric evaluation of interactive web-based text mining technologies.These tasks have resulted in valuable resources for developing and evaluating biomedical text mining systems. Overall 53 unique research teams participated across the tracks, representing more than 120 researchers, with some researchers taking part in multiple tracks (CDR: 24 teams, CHEMDNER-patents 22 teams, BioC 9 teams, IAT 6 teams, BEL 5 teams).The BioCreative V evaluation workshop (http://www.biocreative2015.org, 73 registered participants) provided a forum to discuss the results of each track and the participating text-mining systems.

Additionally, there were three panel sessions on (a) text mining for literature curation, (b) crowdsourcing and curation and (c) disease annotation and medical literature. These sessions provided a forum to discuss current trends and limitations as well as future directions for biomedical text mining related to these trending research topics.The 63 workshop proceedings papers of BioCreative V providing descriptions of the evaluation and participating systems are available at http://www.biocreative.org/resources/biocreative-v/proceedings-biocreative5; a special issue of the journal Database is in preparation.

74 Biocuration and Text Mining: Lessons learned from developing an interoperable collaborative biocurator assistant tool for BioGRID.

Rezarta Islamaj Dogan, Sun Kim, Andrew Chatr-Aryamontri, W. John Wilbur and Donald C. Comeau

The purpose of the BioC track in BioCreative V was to create a set of complementary modules that could be seamlessly integrated into a system capable of assisting BioGRID curators. Specifically, the resulting interactive system triaged sentences from full text articles in order to identify text passages reporting mentions and experimental methods for protein-protein and genetic interactions. These sentences were then highlighted in the curation annotation suite. The task required the identification of passages or sentences describing genes/proteins/species involved in the interaction and mentions and/or experimental methods for molecular interactions. Nine teams from all over the world developed one or more modules independently, integrated via BioC, to insure the interoperability of the different systems. The collaborative curation tool task provided several important achievements. 1. A fully operational system, which was achieved in three months of on-line collaboration between the teams. 2. Interoperability. Data was received, produced and exchanged in the BioC format. This simple format avoided many interoperability hurdles. 3. An easy-to-use system. The four participating curators gave positive feedback regarding the user-friendliness and the curation tool in general.
4. Annotated data. A corpus of 120 full text articles is available for the community containing curation-relevant annotations for mentions and experimental methods evidence for protein-protein and genetic interactions. Text mining based tools provide valid support to biocuration only if easy-to-learn and user-friendly. Reaching this goal requires the adoption of new evaluation metrics. The extended, direct interaction between text miners and curators permitted the identification of new questions, challenges, and opportunities for using text mining in manual annotation pipelines.

75 Strategies towards digital and semi-automated curation in RegulonDB

Fabio Rinaldi, Socorro Gama, Hilda Solano Lira, Alejandra Lopez-Fuentes, Oscar Lithgow and Julio Collado-Vides

Life science databases play a crucial role in the organization ofscientific knowledge in the life science domain. The vastity andcomplexity of the life sciences require the presence of ``knowledgebrokers" who identify, organize and structure information derivedfrom experimental results and extracted from the literature. Thisrole is played by database curators, who act as intermediarybetween the producers of the knowledge (experimental scientist) andits consumers.This important role requires highly-skilled individuals who have thebiological expertise needed to recognize the crucial information thathas to be inserted in a specific database. The complexity of the taskand the specific competences required cannot cannot be fully replacedby automated systems, if the aim is to obtain the same quality ofresults.Nevertheless, it is becoming increasingly clear that this traditionalapproach cannot possibly cope with the deluge of new information beingcreated by experimental scientists. PubMed, the reference repositoryof biomedical literature, at present contains more than 25 Millionbibliographical entries, and it grows at a rate of about twopublications per minute.Considering that only a very small subset of the information containedin a paper is actually required by a specific database, it appears asa waste of time and resources that curators often have to read fullpapers in order to find those items that they need.We propose automated strategies that help curators quickly locate thatcrucial information, and provide tools that support them intransposing this information from the paper to the database. Through acombination of text mining and a user-friendly interface, based ontext filters and partially pre-filled forms, curators canconspicuously enhance their efficiency, thus giving them theopportunities to process larger amount of documents, without losingthe quality of the traditional manual curation approach.The ultimate goal is to obtain a high throughput curation with thesame quality as manual curation, or, when such level of quality cannotbe reached, at least be able to provide a quantifiable measure of thedifference in quality.The work presented here is part of an NIH-sponsored collaborativeproject aimed at improving the curation process of the RegulonDBdatabase. RegulonDB is the primary database on transcriptionalregulation in Escherichia coli K-12, containing knowledge manuallycurated from original scientific publications, complemented with highthroughput datasets and comprehensive computational predictions.In this paper we describe the integration of text mining technologiesin the curation pipeline of the RegulonDB database, and discuss howthe process can enhance the productivity of the curators. Among theinnovations that we propose, we describe in particular:- the integration of ""text filters"" in the curation interface, enabling curators to focus on the parts of the text most likelyto yield the information that they are looking for- partially self-filling forms, compiled by the system usinginformation from the paper, leaving to the curator the decisionwhether to accept or modify- a novel semantic linking capability, which enable curators toexplore related information in other papers

76 Retrieving Biomedical Literature: An Open Source Search Engine Based on Open Access Resources

Hayda Almeida, Ludovic Jean-Louis and Marie-Jean Meurs

The retrieval of biomedical literature is a critical task for scientific researchers and health care practitioners. Open scientific literature databases contain a massive amount of data, which is extensively used to support various research activities in life sciences. Lots of research efforts have been made towards improving the retrieval of bioliterature, but the task is still challenging. PubMed and PubMed Central (PMC) are scientific literature databases maintained by the U.S. National Library of Medicine. As of February 2016, PubMed holds over 25 million records, allowing users to search the content of article abstracts, while PMC holds over 3.7 million of free full-text articles. When utilizing databases such as PubMed and PMC to retrieve relevant information, researchers generally need to express their search needs using a specific query language. This makes the task difficult for users not experienced with query languages, and can compromise the knowledge discovery process.In this work, we present an open source search engine that aims to address two different aspects related to the retrieval of biomedical literature: improve the content access offered by PubMed or PMC, and facilitate the query formulation for users by processing queries in natural language. The system is composed of two modules: the indexation module and the complex query module. Based on the search platform Solr/Lucene, the indexation module generates the inverted index of the dataset, representing all documents using relevant content found in the article content (titles, abstract, body, keywords, references, etc.). The complex query module handles complex user queries, which are processed according to different query types. For each type, a specific search strategy is applied to better meet the user needs. In addition, query terms can be expanded using UMLS concepts.Our search engine was created based on the open-access scientific literature made available by the PubMed Baseline Database (BD), and the PMC Open Access (OA) Subset repository. A total of 25,403,053 articles from these sources was indexed as of February 2016. Information retrieval systems are often evaluated using reference judgments or pseudo-judgments. Here we proposed an evaluation method based on pseudo-judgments, and sets of annotated queries. Our evaluation dataset is composed of query-document sets manually annotated by curators working on the mycoCLAP database. The dataset utilized for preliminary evaluation has 19 query-document relations. From the total, 9 queries have a correct response document mapped to a PMC OA entry (full text article). The other 10 have a correct response document mapped to a PubMed BD entry (article abstract). For each query, we analyzed the first 20 ranked documents, and computed a Mean Reciprocal Rank (MRR) score for the correct response document, considering the position where it was found in the search result list. The MRR score over 0.5 indicates that the system retrieved the correct response document in first or second positions for more than half of the requests. Our work currently focuses on improving the retrieval of full-text documents.

77 Collecting Text Mining Resources for GlycoBiology - an Application Case of PubAnnotation and PubDictionaries

Jin-Dong Kim, Toshihide Shikanai, Shujiro Okuda and Shin Kawano

PubAnnotation is a literature annotation repository, to which annotations made to scientific literature is collected. Particularly, its current primary target is life science literature: PubMed articles and PMC Open Access articles. So far, annotation data sets produced by many different groups have been collected and integrated. Examples include entity annotations and relation annotations. Entity types vary from proteins to species or diseases. Some are linguistic annotations like syntactic parses or coreferences. Some are automatic annotations and some are fully manual ones. Thanks to contribution from many groups, those data sets are now accessible in an integrative way.PubAnnotation also features a function to pull annotations from annotation servers. PubDictionaries is an example of annotation server. In fact, PubDictionaries is a repository for dictionaries. User-generated dictionaries, e.g., an Excel file with a collection of protein names and their UniProt IDs, are collected to PubDictionaries. Anyone who has such a dictionary can upload its CSV dump file to PubDictionaries. Then, a REST web service for text annotation based on the dictionary is immediately enabled. Using PubDictionaries, a user can quickly produce annotations based on dictionaries of his/her interest, and using PubAnnotation, he/she can easily check if there are already existing relevant annotations.As an application case, we compiled text mining resources for GlycoBiology, which we call GlycoTM. The GlycoTM collection is now only in a preliminary state. However, it demonstrates how such a collection can be produced using PubDictionareis and PubAnnotation. We will keep developing the GlycoTM collection as an open resource.For GlycoTM, 8 dictionaries are created for glycobiology from 4 databases and ontologies.Following is descriptions of the databases. (1) GlycoEpitope is a database to integrate carbohydrate antigens and their antibodies, as well as the related information such as glycoproteins, glycolipids, enzymes, tissues and diseases. (2) PACDB (Pathogen Adherence to Carbohydrate Database) contains literature-reported and experimentally obtained information of glycans and pathogens (binding and unbinding) and offers ontology-systemized data. It is well known that pathogens recognize glycans. The registered data were reconstructed with ontology implementation. The ontologies were named as PAConto. (3) GDGDB (Glyco-Disease Genes DataBase) contains information of disease induced by alteration of glycosyltransferase genes and glycosidase genes. Ontology-systematized data of glycan metabolism and clinical condition are also maintained. (4) cGGDB (Caenorhabditis elegans GlycoGene Database) is a database for C. elegans glycogenes. This database was designed so that researchers and students in medical biology can easily understand how glycogenes related to human disease are acting in a model organism, C. Elegans.Based on the dictionaries, text annotation collection is produced. Annotation is made to 2,931 PubMed articles from GlycoBiology (Oxford Journal). 5 annotation data sets are produced based on the 5 databases described above. Besides, annotations based on GO, FMA, and ICD10 lexicon or ontologies are also produced as supporting resources. Those annotations are all accessible through REST API of PubAnnotation. An example excerpt can be checked through this URL: http://pubannotation.org/docs/sourcedb/PubMed/sourceid/22459802/spans/689-836/annotations/visualize

78 Identifying Chemical-Disease Relationships Using Multiple Kernel Learning Method

Yan Liu, Yueping Sun, Li Hou and Jiao Li

Relations between chemicals and diseases (Chemical-Disease Relations or CDRs) play critical roles in drug discovery (toxicity), biocuration, pharmacovigilance, etc. Due to the high cost of manual curation and rapid growth of the biomedical literature, several attempts have been made to extract CDRs using automatic systems. In this study, we proposed a kernel learning method to identify CDRs from PubMed. Kernel based learning algorithms have gained more and more popularity in the machine learning community for its solid foundation and promising performance. Compared with single kernel function, multiple kernel function has been proved to have better performance. However, multiple kernel function still lacks of systematic research. In this study, we extracted semantic relations from text using machine learning method which based on multiple kernel function. First of all, we constructed different kernel functions according to different size text object, so as to reflect different semantic features of the corresponding text. Then, we constructed multiple kernel learning framework by combining two or more single kernel function, and achieving optimal extraction of semantic relations. Finally, we verified the effectiveness of the proposed algorithm by applying it to the BioCreative V corpus released in 2015, and give the comprehensive evaluation in accordance with international standards. The result shows that our algorithm which based on multiple kernel function has better efficiency and accuracy.

79 Extracting value from biomedical literature

Parthiban Srinivasan

The research community is now flooded with scientific literature, with thousands of journals and over 20 million abstracts in PubMed. Somewhere in this information lie the answers to questions not only for scientific research but also for business research. A lot of times, a market researcher starts with a question, then collects data and answers the question. But now we can start with public data. Then we figure out a new, useful and valuable question we can ask and answer. Customers want digestible information - everything relevant-not hundreds of journal articles to read. In this talk, we will present case studies on how we used the ontologies and disambiguation techniques to address the needs for business analytics in the pharmaceutical research. The results will be presented in the context of identification of key opinion leaders.

80 Updating and Extending the Concept Annotations of the CRAFT Corpus

Michael Bada, Nicole Vasilevsky, Melissa Haendel and Lawrence Hunter

With the ever-rising amount of biomedical literature, it is increasingly difficult for scientists to keep up with the published work in their fields of research, much less related ones. The use of natural language processing (NLP) tools can make the literature more accessible by aiding concept recognition and information extraction. As NLP-based approaches have been increasingly used for biocuration, so too have biomedical ontologies, whose use enables semantic integration across disparate curated resources, and millions of biomedical entities have been annotated with them. Particularly important are the Open Biomedical Ontologies (OBOs), a set of open, orthogonal, interoperable ontologies formally representing knowledge over a wide range of biology, medicine, and related disciplines.Manually annotated document corpora have become critical gold-standard resources for the training and testing of biomedical NLP systems. This was the motivation for the creation of the Colorado Richly Annotated Full-Text (CRAFT) Corpus, a collection of 97 full-length, open-access journal articles from the biomedical literature. Within these articles, each mention of the concepts explicitly represented in eight prominent OBOs has been annotated, resulting in gold-standard markup of genes and gene products, chemicals and molecular entities, biomacromolecular sequence features, cells and cellular and extracellular components and locations, organisms, biological processes and molecular functionalities. With these ~100,000 concept annotations among the ~800,000 words in the 67 articles of the 1.0 release, it is one of the largest gold-standard biomedical semantically annotated corpora. In addition to this substantial conceptual markup, the corpus is fully annotated along a number of syntactic and other axes, notably by sentence segmentation, tokenization, part-of-speech tagging, syntactic parsing, text formatting, and document sectioning.In the several years since the initial release of the CRAFT Corpus, in addition to efforts within our group and in collaboration with others, including the first comprehensive gold-standard evaluation of current prominent concept-recognition systems, it has already been used in multiple external projects to drive development of higher-performing systems. Here we present our continuing work on the corpus along several fronts. First, to keep the corpus relevant, we are updating the concept annotations using newer versions of the ontologies already used to mark up the articles, removing annotations of obsoleted classes and editing previous annotations or creating new annotations of newly added classes. Additionally, to extend the domain of annotated concept types, we are also marking up mentions of concepts using the Molecular Process Ontology (for types of chemical processes) and the Uberon Anatomy Ontology (for anatomical components and life-cycle stages). Finally, to capture even more content, we are generating new annotations for roots of prefixed/suffixed words as well as annotations made with extension classes we have created. We will present updated annotation counts and interannotator agreement statistics for these continuing efforts as well as future plans. All of this work is designed to further increase the potential of the CRAFT Corpus to significantly advance biomedical text mining by providing a high-quality gold standard for NLP systems.

81 A Curation Pipeline for Bio-derived Chemical Feedstocks

George Demetriou, Warren Read, Martyn Fletcher, Noel Ruddock, Goran Nenadic, Tom Jackson, Robert Stevens and Jerry Winter

The BioHub Information and Knowledge Management System (IKMS)aims to support the process of identifying chemicalingredients that can be sourced from sustainable biomass asalternatives to those from non-renewable resources suchas fossil oils or earth minerals.The curation of chemical data in BioHub is performed in three stages:(i) A text mining stage whose aim is to mine facts about chemicals directlyfrom the scientific literature (journal papers), patents and laboratory reports).(ii) An 'assertion generation' stage where 'assertions' (i.e. factual statements about feedstocks and chemicals) are selected as candidates for curation by querying the text mining results.(iii) A curation stage that allows curators to browse, edit,validate and store the assertions to the system's final data store.The text analytics system of BioHub is developed within the GATEplatform. It processes documents in order to extractrelevant information about chemicals such as the feedstock streamsfrom which they are derived, their physical and chemical properties, possible transformations applied to them etc. This information isexported to OWL data stores enriched with links to the parts of thetexts from which it comes.Curation is not performed directly on the text mining results but on the output ofsemantic queries applied to them. Assertions are generated on-the-fly via SPARQL-ingthe OWL output and are transformed to JSON objects ready to be parsedand presented in tabular format on the BioHub curation User Interface(UI). The conceptual structure of an assertion is pre-specified.An example of an assertion type is ""feedstock-has-chemical-with-proportion"".This is derived from two binary relations (triples) identified in the text mining stage i.e. ""feedstock-has-chemical"" and ""chemical-has proportion"", both ofwhich may have been extracted from the same part of the text (usually asentence) and have been linked together by querying the text mining data.The main goal of the curation UI is to populate the BioHub datarepository with validated assertions. Secondary goals are to aidperformance evaluation and provide curated data for system trainingand refinement. Its design has been influenced by considerations of anintuitive interface aiming at rapid curation rather thanconsiderations of a full-fledged annotation editor which mightrequire much more user intervention and time resources for curatingfull text documents. The UI is employable by a Web-based client withclient-side services for input data and server-side facilities for storing thefinal results. Its editing engine includes capabilities such ascontext-enabled curation (allows access to the textsentences, paragraphs or full text), content-editable fields forentities, grouping/sorting of assertions per various facets, logging, etc.The results of the curation stage are stored to the BioHub datarepository and are used to support subsequent stages in the IKMS,such as the selection of ingredients based on functionalcharacteristics and the computational optimisation of chemical pathways.The curation architecture of the BioHub IKMS demonstrates how textmining and semantic web technologies can be integratedwithin a distributed, goal-oriented curation infrastructure to facilitate thesemi-automated development of knowledge bases in the chemical domain.

82 Outreach activities at FlyBase

Alix Rey, Laura Ponting, Gary Grumbling, Jim Thurmond, Jose-Maria Urbano, Gillian Millburn, Steven Marygold and Nick Brown

FlyBase has an active and diverse outreach program to engage with our user community. For example, the FlyBase Community Advisory Group (FCAG) comprises over 550 FlyBase users from around the world and provides essential feedback on new features and changes to FlyBase through regular surveys. We have also implemented the online "Fast-Track Your Paper" tool to facilitate community curation, with over 50% of authors routinely associating genes with their publications in FlyBase and highlighting data types requiring a deeper curation.More recently, we have starting making video tutorials as a means to answer common queries and help people get the most of out of FlyBase. Topics covered so far include 'How to find all data related to a gene', 'How to generate an excel file of all alleles of a gene' and 'How to cite FlyBase'. To produce a video, a script is first written, then a screen recording is captured and a voice is added to it. The videos are available on the newly created FlyBase YouTube channel, FlyBase TV. Subsequent videos will focus on each of the various tools in FlyBase.

83 Predicting Structured Metadata from Unstructured Metadata

Lisa Posch, Maryam Panahiazar, Michel Dumontier and Olivier Gevaert

Enormous amounts of biomedical data have been and are being produced by investigators all over the world. However, one crucial and limiting factor in data reuse is accurate, structured and complete description of the data, or data about the data - defined as metadata. We propose a framework to predict structured metadata terms from unstructured metadata for improving quality and quantity of metadata, using the GEO microarray database. Our framework consists of a Latent Dirichlet Allocation model (LDA) to reduce the dimensionality of the unstructured data, in combination with a supervised classifier. We compared support vector machines and decision trees with the majority classifier as baseline. Our results on the GEO database show that structured metadata terms can be accurately predicted. This is a promising approach for metadata prediction that is likely to be applicable to other datasets and has implications for researchers and practitioners interested in biomedical metadata curation and metadata prediction.

84 Automated Detection of Discourse Segment and Experimental Types from Cancer Pathway Primary Research Articles

Gully Burns, Pradeep Dasigi, Anita de Waard and Eduard Hovy

Automated machine-reading biocuration systems typically use sentence-by-sentence information extraction to construct meaning representations for use by curators. This does not directly reflect the typical discourse structure used by scientists to construct an argument from the experimental data available within a paper, and is therefore less likely to correspond to representations typically used in biomedical informatics systems. In this study, we develop Natural Language Processing (NLP) methods to locate, extract and classify the individual passages of text from papers' results sections that refer to experimental data. In our domain of interest (molecular biology studies of cancer signal transduction pathways), individual papers may have as many as 30-50 individual experiments describing a variety of findings, which authors use to base their research narrative on. The early results of the work described here are concerned with classifying discourse segments in these texts into seven categories (fact, hypothesis, problem, goal, method, result, implication) which collectively describe the essential building blocks of scientific discourse to (A) provide context for each experiment, (B) report experimental details and (C) explain the data's meaning. We next evaluated text passages from articles that had been curated in two molecular biology databases (the Pathway Logic Datum repository and the Molecular Interaction 'MINT' database) linking individual experiments in papers to the type of assay used (coprecipitation, phosphorylation, translocation, etc.). We used simple supervised machine learning techniques on text passages containing unambiguous references to experiments to yield baseline F-Scores of 0.59 for MINT and 0.63 for Pathway Logic. These results support the notion that targeting information extraction methods to experimental results could provide accurate, automated methods for biocuration.

85 Manual Curation in the Conserved Domain Database

Noreen Gonzales, Farideh Chitsaz, Myra Derbyshire, Lewis Geer, Marc Gwadz, Lianyi Han, Jane He, David Hurwitz, Christopher Lanczycki, Fu Lu, Gabriele Marchler, James Song, Narmada Thanki, Josie Wang, Roxanne Yamashita, Chanjuan Zheng, Steve Bryant and Aron Marchler-Bauer

The Conserved Domain Database (CDD) is a collection of multiple sequence alignments that represent ancient conserved domains. One part of the CDD resource is a mirror of publicly available domain model collections, including Pfam and TIGRFAMs, among others. These may be used as starting points for manually-curated conserved domain models (accessions with a "cd" prefix) arranged in a hierarchical structure to reflect evolutionary diversification of ancient protein domain families. Most curated models contain annotation of features that are conserved across the domain family, supported by evidence obtained from 3D structures as well as the published literature. Curated domain family models are also created de-novo for previously uncharacterized families, often identified via novel 3D structures with no conserved domain annotation. Hierarchical classification and curation of protein domains, using our in-house tools CDTree (hierarchy viewer) and Cn3D (structure viewer and multiple alignment editor), have been the focus of our manual curation efforts. In addition, we develop structural motif models (accessions with an "sd" prefix) to represent protein sequence segments such as short repeats, coiled coils, and transmembrane regions. We also manually validate superfamily clusters (accessions with a "cl" prefix), formed by an automated clustering procedure as sets of conserved domain models that generate overlapping annotation on the same protein sequences. Superfamily clustering allows the organization of data within CDD in a non-redundant way, as each data source may have its own model for a specific conserved domain. Cluster validation is aided by using Cytoscape as a visualization tool for the degree of overlap between conserved domain models. More recently, our manual curation efforts are focused on providing functional labels for domain architectures, using an in-house procedure called SPARCLE (""Specific ARChitecture Labeling Engine""). While we are able to assign functional labels to a large fraction of proteins, we have also identified areas of insufficient coverage and resolution of the current protein domain models that comprise CDD. In this poster, we will discuss all aspects of manual curation in CDD. The need for manual curation work always exceeds available resources and we hope to automate hierarchical classifications to some degree in the near future.AcknowledgementThis research was supported [in part] by the Intramural Research Program of the National Library of Medicine, NIH.

86 Curation of reference malaria parasite genomes

Ulrike Boehme, Thomas Dan Otto, Mandy Sanders, Chris Newbold and Matthew Berriman

The genomes of seven malaria parasite species (Plasmodium spp) are currently being curated, including those of the rodent-malaria parasites, P. chabaudi, P. yoelii and P. berghei; the human-infective species, P. falciparum, P. vivax and P. knowlesi; and the chimpanzee parasite P. reichenowi. Thousands of additional genomes are being sequenced from clinically isolated parasites from across the globe to study the evolving genetics of parasite populations. In addition, draft genomes of additional species are being used to understand the structure and evolution of Plasmodium genomes.

Manual curation of all of these data would be impossible. Therefore we are focusing curation activities on one genotype per reference species. This enables the malaria community to use these reference genomes to look for manually curated GO terms and products and transfer them to other sequenced isolates that are not curated. We have established a workflow for manual curation. The annotation tool Artemis is used to read and write directly to a Chado relational database underlying GeneDB (http://www.genedb.org). An annotation-transfer tool has been implemented in Artemis to transfer annotation between features within the same Chado database. GeneDB houses curated Plasmodium reference genomes and is being updated daily. As part of a collaborative effort with PlasmoDB (http://www.plasmodb.org) every few months the annotated and curated genomes are sent from GeneDB to PlasmoDB to be integrated with a wide variety of functional genomics data sets. PlasmoDB enables the community also to search non-curated genomes that are not being updated. A banner with a direct link to the GeneDB gene record page has been implemented to inform the community of changes in the annotation.

87 PROSITE, a database for domain and site detection and annotation

Christian J. A. Sigrist, Edouard de Castro, Beatrice A. Cuche, Delphine Baratin, Thierry Schuepbach, Sebastien Moretti, Marco Pagni, Sylvain Poux, Nicole Redaschi, Alan Bridge, Lydie Bougueleret and Ioannis Xenarios

PROSITE is a resource for the identification and annotation of conserved regions in protein sequences. These regions are identified using two types of signatures: generalized profiles (weight matrices) that describe protein families and modular protein domains and patterns (regular expressions) that describe short sequence motifs often corresponding to functionally or structurally important residues. PROSITE signatures are linked to annotation rules, or ProRules, which define protein sequence annotations (such as active site and ligand-binding residues) and the conditions under which they apply (for example requiring specific amino acid residues). PROSITE signatures, together with ProRule, are used for the annotation of domains and features of UniProtKB/Swiss-Prot entries. The latest version of PROSITE (release 20.122, of 13 January 2016) contains 1309 patterns, 1145 profiles and 1145 ProRules and is accessible at: http://prosite.expasy.org/prosite.html. The ScanProsite tool (http://prosite.expasy.org/scanprosite/) allows users to search protein sequences against all PROSITE signatures, and to search for matches to defined PROSITE signatures in the UniProtKB and PDB databases. Individual protein sequences and whole proteomes can be subjected to repeated scans with the benefits of the PROSITE graphical view of the results and the application of ProRule for a more precise prediction.

88 Prediction of Metabolic Pathway Involvement in Prokaryotic UniProtKB Data by Association Rule Mining

Imane Boudellioua, Rabie Saidi, Robert Hoehndorf, Maria J. Martin and Victor Solovyev

The widening gap between known proteins and their functions has encouraged the development of methods to automatically infer annotations. Functional annotation of proteins is expected to meet the conflicting requirements of providing comprehensive information while avoiding erroneous functional assignments. This trade-off imposes a great challenge in designing intelligent automatic annotations systems. In the scope of this work, we tackle the problem of UniProtKB automatic functional annotation of prokaryotic pathways. We suggest that association rule mining could be used effectively as a computational method for pathway prediction. Here, we introduce ARBA, an Association-Rule-Based Annotator that can be used to enhance the quality of automatically generated annotations as well as annotating proteins with unknown function. ARBA utilizes data from UniProtKB/Swiss-Prot and uses InterPro signatures and organism taxonomy as attributes of predict metabolic pathways associated with each protein entry. With respect to certain quality measures, we find all rules which would define significant relationships between attributes and pathway annotations in UniProtKB/Swiss-Prot entries. The set of extracted rules represent the comprehensive knowledge which could explain protein pathway involvement. However, these rules comprise redundant information and their high number makes it infeasible to apply them on large sets of data such as UniProtKB/TrEMBL. To address this issue, ARBA puts these rules into a fast competition process based on two concepts, namely dominance and comparability. Rules are then considerably reduced in number and aggregated with respect to the predicted pathways. The resulting knowledge represents concise prediction models that assign pathway involvement to UniProtKB entries. We carried out an evaluation study of our system's performance using semantic similarity and cross-validation technique on UniProtKB prokaryotic entries to demonstrate the performance, capability and robustness of our approach. We found that we achieved a very high accuracy of pathway identification with an F1-measure of 0.982 and AUC of 0.987. Moreover, our prediction models were applied on 6.2 million UniProtKB/TrEMBL reference proteome entries of prokaryotes. As results, (663,724) entries were covered, where (436,510) of them lacked any previous pathway annotations. Observing the annotation coverage of this set of entries in comparison to other main automatic annotation systems present in UniProtKB/TrEMBL which are, SAAS and UniRule (which includes Rule-base and HAMAP-Rule), we found out that ARBA significantly surpassed the other ones in terms of the number of entries covered. As stated earlier, ARBA annotated (663,724) entries where HAMAP-Rule, SAAS, and Rule-base annotated only (229,402), (205,097) and (93,613) entries respectively. On the other hand, analyzing the entries annotation, we found that (786,819) predictions were generated by ARBA where the majority of these predictions, (516,042), touched entries that have no previous pathway annotation. Moreover, (237,784) predictions were found to be identical to the annotations proposed by other systems which enforce the reliability of our systems' predictions.A Java Archive (JAR) package for applying the prediction models on various UniProtKB/TrEMBL prokaryotic entries is available at: http://www.ebi.ac.uk/~rsaidi/arba/The link also contains the list of prediction models and graphical reports illustrating the system's performance on some prokaryotic organisms in UniProtKB/TrEMBL.

89 UNCOVERING HUMAN TRANSCRIPTIONAL REGULATION THROUGH THE SEQUENCE INFORMATION

Neven Sumonja, Vladimir Perovic and Nevena Veljkovic

A large part of post-genomic research is focused on the analysis of protein-protein interactions (PPIs) being central to all biological processes. Inferring PPIs involved in human transcriptional regulation (TR) is of particular interest as they are often deregulated in complex diseases and may represent valuable pharmaceutical targets. We devised a method to analyze and predict these interactions based on sequence information only aiming to evade limitations imposed by dispersed auxiliary information, such as localization, structural and expression data. A new predictor incorporates information on the pseudo-amino acid composition of features that dominate PPIs. Besides the electrostatic and hydrophobic features, it incorporates the electron-ion interaction potential (EIIP) a descriptor of long-range interaction properties that contribute to protein binding specificity through long-range recognition between partners. Based on a dataset that was compiled from HIPPIE (Human Integrated Protein-Protein Interaction rEference), a random forest model was constructed with an average value of accuracy of 80.41% and AUC 0.88 for independent test sets. Compared with previous studies, our approach outperformed other models in predictive performance and algorithmic efficiency and will, therefore, facilitate the understanding of the complex cellular behaviors and organizing of large-scale data into models of cellular signaling and regulatory machinery.

90 Rapid and Efficient Method for Phylogenetic Analyses Based on Digital Signal Processing Techniques

Vladimir Perovic, Veljko Veljkovic, Sanja Glisic and Nevena Veljkovic

As sequencing technologies continue to drop in price and increase in throughput, new challenges regarding processing the vast datasets for a huge number of genes and identifying an optimal analytical methodology emerge. The informational spectrum method (ISM) is a sequence analysis approach that relies on the Fast Fourier Transform (FFT) combined with decoding of the protein sequence via amino acid physicochemical properties that transforms the sequence into an Informational Spectrum (IS). Starting from the IS of the protein, we developed new protein distance measures and a novel phylogenetic algorithm ISTREE that has been found to overcome some drawbacks of classical phylogenetic approaches, particularly those related to sensitivity to a single mutation and deletion and as well as to the position of the mutation. Given that ISTREE is based on FFT and does not require multiple sequence alignment (MSA), it is a fast method for evolutionary analyses of large sets of protein sequences, compared to other standard phylogenetic algorithms. We used ISTREE to study the functional evolution of the hemagglutinin subunit 1 protein (HA1), in an effort to better understand the viral determinants that facilitate human infections of the highly pathogenic avian influenza (HPAI) A subtype H5N1 virus. The mutations that increase HPAIV propensity for human-to-human transmission were identified and the predictions were confirmed in vitro.

91 Re-annotation of the rice genome (Oryza Sativa Japonica) based on RNA-Seq data

Lili Hao, Jian Sang, Lin Xia and Zhang Zhang

Rice is one of the most important staple food for a large portion of the world's population and also a key model organism for cereal crops due to its great agricultural importance. In order to provide a precise reference genome in aid of extensive rice-related studies, it is desirable to keep annotating the rice genome by integrating large quantities of high-throughput omics data. Here, we present a re-annotation release of the Oryza Sativa Japonica genome (BIGD-IC4R-1.0), for the first time, based on more than 700 publicly available high-quality RNA-Seq datasets (~7.4 Terabyte) along with annotations contributed from NCBI, EBI and UniProt, thereby providing substantial improvements over the previous version MSU Rice Genome Annotation Project Release 7.0 (MSU7.0; released on Feb 6, 2013). Our near-final release of BIGD-IC4R-1.0 consists of 57,905 protein-coding genes, among which 2,259 novel genes are identified for the first time, and the structural annotations of a total of 20,682 genes have been updated based on the previous Version MSU7.0. Moreover, the number of genes in BIGD-IC4R-1.0 with splice variants is significantly increased compared with MSU7.0. In addition, 11,841 long ncRNAs were identified from 658,655 assembled transcripts. BIGD-IC4R-1.0, an updated version for rice genome re-annotation that is, for the first time, based on large-scale RNA-Seq data analysis, has revised hundreds of inaccurate gene models and provided a number of alternatively spliced isoforms as well as long ncRNAs, which thus would be of critical importance for significantly promoting functional studies in rice as well as other plants.

92 Protein Structures and their features in UniProt

Nidhi Tyagi, Guoying Qi, Maria-Jesus Martin, Claire O'donovan and Uniprot Consortium

Annotation of proteins based on structure-based analyses is an integral component of the UniProt Knowledgebase (UniProtKB). There are over 100,000 experimentally determined 3-dimensional structures of proteins deposited in the Protein Data Bank. UniProt works closely with the Protein Databank in Europe (PDBe) to map these 3D structural entries to the corresponding UniProtKB accessions accurately and coherently based on comprehensive sequence and structure-based analyses, to ensure that there is a UniProtKB record for each relevant PDB record and to import additional data such as ligand-binding sites from PDB to UniProtKB. SIFTS (Structure Integration with Function, Taxonomy and Sequences), which is a collaboration between the Protein Data Bank in Europe (PDBe) and UniProt, facilitates the link between the structural and sequence features of proteins by providing correspondence at the level of amino acid residues. A process combining manual and automated processes for maintaining up-to-date cross-reference information has been developed and is carried out for every weekly PDB release. Various criteria are considered to cross-reference PDB and UniProtKB entries such as a) High sequence identity (>90%) b) Exact taxonomic match (at the level of species, subspecies and specific strains for lower organisms) (c) Mapping to curated SwissProt entry (if exists) (d) Mapping to proteins from Reference/Complete proteome (e) mapping to the longest protein sequences. Some cases are inspected manually by UniProt using a dedicated curation interface to ensure accurate cross-referencing. These cases include short peptides, chimeras, synthetic constructs and De novo designed polymers. The SIFTS initiative also provides up to date cross referencing of structural entries to literature (PubMed), taxonomy (NCBI), Enzyme database (IntEnz), Gene Ontology annotations (GO), protein family classification databases (InterPro, Pfam, SCOP and CATH) .In addition to maintaining accurate mappings between UniProtKB and PDB, a pipeline has been developed to automatically import data from PDB to enhance the unreviewed records in UniProtKB/TrEMBL. This includes details of residues involved in the binding of biologically relevant molecules including nucleotides, metals, drugs, carbohydrates and post-translational modifications and greatly improves the biological content of these records.To date, UniProt has successfully completed the non-trivial and labour intensive exercise of cross referencing ~250,000 polypeptide chains and 102,417 PDB entries (out 115,306 entries processed by PDBe). All this work enables non-expert users to see protein entries in the light of relevant biological context such as metabolic pathways, genetic information, molecular functions, conserved motifs and interactions etc. Protein structural information in UniProt serves as a vital dataset for various academic and biomedical research projects.

93 UniRule - Increasing Annotation Depth of Unreviewed Protein Entries in UniProtKB

Klemens Pichler, Ricardo Antunes, Mark Bingley, Emma Hatton-Ellis, Alistair MacDougall, Maria Martin, Diego Poggioli, Sangya Pundir, Alexandre Renaux, Vladimir Volynkin, Hermann Zellner, Cecilia Arighi, John S. Garavelli, Kati Laiho, C.R. Vinayaka, Qinghua Wang, Lai-Su Yeh, Delphine Baratin, Alan Bridge, Edouard de Castro, Ivo Pedruzzi, Nicole Redaschi, Catherine Rivoire, Claire O'Donovan and Uniprot Consortium

The mission of UniProt is to provide a comprehensive and thoroughly annotated protein resource to the scientific community, most notably through the UniProt Knowledgebase (UniProtKB). Within UniProtKB, the reviewed section (Swiss-Prot) contains high quality, manually curated, richly-annotated protein records. In contrast, the unreviewed section (TrEMBL) which makes up 99% of UniProtKB, depends for its annotation on automatically extracted experimental data from 3D structures, links to other databases and rule-based annotation. The use of rule-based annotation is necessary because there is no experimental data available for the majority of the unreviewed protein sequences. This makes inference of function by similarity/homology the only option for annotation. UniRule is a rule-based annotation system leveraging the expert-curated data in reviewed UniProtKB to increase the depth of annotation in unreviewed entries. Currently the UniRule system contains over 4,500 rules, which provide annotation for approximately 28% of unreviewed entries. Rules are a formalized way of expressing an association between conditions, which have to be met, and annotations, which are then propagated. InterPro signatures, predictive models for the functional classification of protein sequences, and taxonomic constraints are the fundamental conditions but others are used, too. Annotation types used by UniRule allow the complete functional annotation of a protein sequence, including nomenclature, catalytic activity, Gene Ontology (GO) terms and sequence features such as transmembrane domains. Data provenance is documented using Evidence Ontology tags. A key feature of the UniRule curation tool is a statistical system which allows curators to evaluate their rules against the reviewed entries, to make sure rules are as accurate as possible. This quality control system also allows curators to re-evaluate and update old rules at every release, ensuring that the propagated annotation in the unreviewed entries is kept up to date. A dedicated space on the uniprot.org website has recently been created to allow users to view and explore UniRule. All aspects of the UniRule system and the latest developments will be explained at the conference.

94 HAVANA manual annotation of vertebrate genomes

Deepa Manthravadi, Ruth Bennett, Alexandra Bignell, Gloria Despacio-Reyes, Sarah Donaldson, Adam Frankish, James Gilbert, Michael Gray, Ed Griffiths, Gemma Guest, Matt Hardy, Toby Hunt, Mike Kay, Jane Loveland, Jonathan Mudge, Gaurab Mukherjee, Charles Steward, Marie-Marthe Suner, Mark Thomas and Jennifer Harrow

The HAVANA group at the Wellcome Trust Sanger Institute manually annotate the gene content of vertebrate genomes. We annotate using the in-house Zmap viewer interface, which transfers models to the Otter software for storage and processing and is publicly available. We aim to produce complete genesets for Human, Mouse and Zebrafish which is done in a clone by clone or gene targeted manner, annotating all protein coding genes, non-coding RNAs and pseudogenes. We are also engaged in community annotation projects chiefly for pig and rat, where collaborators request loci or areas of interest for HAVANA annotation. HAVANA classify transcripts according to functional 'biotypes'. We have numerous coding and non-coding biotypes, reflecting our confidence in the annotation of the sequences (known, putative, nonsense mediated decay), as well as a sophisticated system of pseudogene classification. We have incorporated next generation datasets such RNAseq, CAGE and PolyASeq data into our annotation workflow. These datasets are particularly important for transcriptomes that lack coverage from traditional datasets e.g. zebrafish. However, even our human geneset remains a work in progress, and we have recently begun to use long-read RNAseq PacBio data as well as the synthetic long reads produced by Tilgner et al. These are used to identify additional alternatively spliced transcripts, to complete existing partial models and even to find new loci. We are also using PhyloCSF to help identify additional coding regions in human and mouse, especially those that may already have been missed during our first-pass annotation. In addition, we are collaborating with the proteomics group at Sanger to identify peptides produced from mass spectrometry that support the coding potential of transcripts previously annotated as non-coding.All of our data is publicly available from the VEGA website (www.vega.sanger.ac.uk). The GENCODE genes-sets for human and mouse can also be accessed from the UCSC and Ensembl genome browsers, as well as the GENCODE web portal (www.gencodegenes.org)

95 Large-scale gene functional inference through orthology

Robert Waterhouse, Felipe Simao, Panagiotis Ioannidis, Evgenia Kriventseva, Evgeny Zdobnov and Mirna Tenan

Orthology delineation is a cornerstone of comparative genomics that offers evolutionarily-informed hypotheses on gene function by identifying "equivalent" genes in different species. The OrthoDB catalog of orthologs, www.orthodb.org [Kriventseva, et al. 2015], represents a comprehensive resource of comparative genomics data to help researchers make the most of their newly-sequenced genomes. The rapid accumulation of sequenced genomes mean that such comparative approaches are becoming ever-more powerful as tools to improve both genome-wide gene structural annotations and large-scale gene functional inferences. Orthology delineation offers a solid foundation from which to begin to interpret characteristic genome biology traits of a species or clade of species, highlighting shared and unique genes that offer clues to understanding species diversity and providing the means to begin to investigate key biological traits, for both large-scale evolutionary biology research and targeted gene and gene family studies. The OrthoDB catalog collates available gene functional information from UniProt, InterPro, GO, OMIM, model organism phenotypes and COG functional categories, as well as providing evolutionary annotations including rates of ortholog sequence divergence, gene copy-number profiles, homology-related sibling groups and gene architectures. These resources enable improved and extended orthology-based gene functional inference in a comparative genomics framework that incorporates the rapidly growing numbers of newly-sequenced genomes. Such approaches are well-established as immensely valuable for gene discovery and characterization, helping to build resources to support biological research. The success of such interpretative analyses relies on the comprehensiveness and accuracy of the input data, making quality assessment an important part of the process of genome sequencing, assembly, and annotation. OrthoDB's sets of

Benchmarking Universal Single-Copy Orthologs, BUSCO [Sim~ao, et al. 2015], provide a rich source of data to assess the quality and completeness of these genome assemblies and their annotations. Orthology-based approaches therefore offer not only a vital means by which to begin to interpret the increasing quantities of genomic data, but also to help prioritize improvements, and to ensure that initial "draft" genomes develop into high-quality resources with evolutionarily-informed gene functional inferences that benefit the entire research community.

96 Hidden in plain sight: The eukaryotically conserved unstudied proteins and a framework for their classification and characterisation

Valerie Wood, Midori Harris and Antonia Lock

Proteins conserved widely among eukaryotes play fundamentally important roles in the shared, basic mechanisms of life. The roles of many broadly conserved proteins remain unknown, however, despite almost a century of genetic and biochemical investigation. Even the recent emergence of genome-wide techniques and the availability of near-complete protein inventories for many intensively studied eukaryotic model species have shed light on the functions of few previously uncharacterised conserved proteins. Because the success of many endeavours in basic and translational research (drug discovery, metabolomics, systems biology), depends critically on comprehensive representation of functions, a more complete understanding of protein components conserved throughout eukaryotes would have far-reaching benefits for biological research in many species. To identify priority targets for experimental investigation, PomBase provides an inventory of fission yeast proteins that are conserved among eukaryotes but whose broad biological roles remain unknown. A broad functional classification of the known proteome using a selection of Gene Ontology biological process categories has revealed correlations with features such as subcellular localization and morphological phenotype. Combining available data from genome-wide phenotype and localization experiments with insights from the functional classification of known proteins facilitates prediction of biological roles, and thereby guides specific experimental characterisation of unknown proteins.

97 Computational Functional Annotation using Hierarchical Orthologous Groups in OMA

Alex Warwick Vesztrocy, Nives Skunca, Adrian Altenhoff and Christophe Dessimoz

Computational methods for Gene Ontology (GO) annotation are essential to keep abreast of the unabated growth of genomic data. In particular, phylogenetic methods provide a compelling framework to propagate gene attributes across related sequences. The Orthologous Matrix (OMA) database propagates GO annotations among orthologous relationships across ~2000 genomes, currently inferring ~80 million GO annotations. Here, two methods shall be presented. The first, currently implemented in OMA, propagates annotations within cliques of orthologous genes. The second propagates terms across Hierarchical Orthologous Groups (HOGs)-nested groups of genes that descend from single ancestral genes-which makes it possible to compare highly diverged and similar species in a consistent manner. Terms get propagated up the hierarchy towards the root, with the belief in them exponentially decaying over each step. At each node in the hierarchy, annotations from the child groups are mixed to decide the overall annotations at that level of the hierarchy. The merits and challenges of these and other functional annotation methods will be discussed, including complications associated to the open-world assumption.

98 Functional labeling of domain architectures with SPARCLE

Aron Marchler-Bauer, Lianyi Han, Christopher Lanczycki, Jane He, Shennan Lu, Farideh Chitsaz, Myra Derbyshire, Noreen Gonzales, Marc Gwadz, Fu Lu, Gabriele Marchler, James Song, Narmada Thanki, Roxanne Yamashita, Chanjuan Zheng, Stephen Bryant and Lewis Geer

SPARCLE, the SPecific ARChitecture Labeling Engine, is a curation interface developed for the Conserved Domain Database team at the National Center for Biotechnology Information. SPARCLE is used to associate conserved domain architectures with suggested protein names and brief functional descriptions or labels, as well as corresponding evidence. Protein names and labels range from generic to very specific, reflecting the status quo of the underlying protein and domain model collections. They may, however, provide concise functional annotations that are easier to interpret than raw representations of domain architecture.This work was supported by the Intramural Research Program of the National Institutes of Health, National Library of Medicine

99 Integration of a semantic framework for high throughput functional genomics with public resources: A comparison of nearly 4000 strains

Jasper Koehorst, Jesse van Dam, Ruben van Heck, Edoardo Saccenti, Vitor A.P. Martins Dos Santos, Maria Suarez Diez and Peter Schaap

High quality genome annotations are fundamental for the understanding of a species. Nowadays automatic annotation pipelines combining ab initio gene prediction with algorithms and homology based approaches for functional annotation are standardly used. It is essential to review these electronically inferred functional annotations. However, data management complications prevent existing pipelines from storing the complete data provenance. This results in annotated genomic features with unknown origin. For comparative genomics this information is essential as different algorithms and methodologies will yield different results. To track annotation with the corresponding provenance and to allow for ease of integration of different data sources we developed an extensible Semantic Annotation Platform for Prokaryotes (SAPP). Due to the fast increasing number of sequenced genomes this platform has been designed to integrate large volumes of genomic data. The platform is modularly designed and it can be extended with new features. SAPP provides SPARQL support for querying and analysing computational results while aggregating heterogeneous data from alternative sources. Phenotypic characterisations were integrated through a collaborative effort fostered by WikiData, interconnecting multiple resources through semantic end-points. We performed an in depth comparative analysis of nearly 4.000 publicly available bacterial genomes. We identified genotype-phenotype associations, pinpointing key features responsible for several bacterial phenotypic traits such as pathogenesis, cell wall characteristics and composition, and environmental requirements. Our results clearly show the potential of semantic technologies to perform large scale comparative genomics

100 Integration of proteomics data into UniProtKB

Emanuele Alpi, Guoying Qi, Alan Da Silva, Benoit Bely, Jie Luo, Maria Martin and Uniprot Consortium

The identification of peptides and proteins in mass spectrometry (MS) based proteomics experiments relies in searching protein sequence databases. Therefore, it is of paramount importance the provision of an up-to-date, stable and complete protein sequence database for a diversity of species.UniProt provides a broad range of reference protein data sets for a large number of species, specifically tailored for an effective coverage of sequence space while maintaining a high quality level of sequence annotations and mappings to the genomics and proteomics information.With respect to publicly available bottom-up proteomics data, UniProt started providing mappings to its reference proteomes from release 2015_03 either in the protein entries and made available on the ftp (ftp.uniprot.org/pub/databases/uniprot/current_release/knowledgebase/proteomics_mapping/) via the download section of the website (www.uniprot.org/downloads).The mappings are recalculated for every UniProt release starting each time from a fresh data retrieval from the collaborating MS proteomics repositories and they contain isoform-specific information which will also be graphically displayed soon on the website through a dedicated feature viewer interface.In addition, the collaborating MS proteomics repositories have been cross-referenced from within UniProt data and website.Since then the mappings have been expanded both in terms of covered species and collaborating MS proteomics repositories. Ongoing collaborations have been established to add other MS proteomics repositories as data providers for the mappings also in order to further expand the range of covered species.Special cases of these collaborations are the ones aimed at global reprocessing of the content of PRIDE (the PRoteomics IDEntifications database) and the ones which will provide data with a specific focus on posttranslational modification (PTM) related studies/datasets.Another very promising collaboration is the one with the Consortium for Top Down Proteomics (CTDP, http://www.topdownproteomics.org/) which has been cross-referenced from within UniProt data and website since release 2016_03.The top-down proteomics data available through the CTDP repository is currently used by UniProt for the development of a dedicated pipeline to annotate back the UniProt entries and publicly provide the corresponding mappings on the ftp.CTDP data include isoform-specific and variant-specific information for whole proteoforms also bearing PTMs.

101 GSA: Genome Sequence Archive

Gsa Project Consortium, Hongxing Lei, Xiangdong Fang and Wenming Zhao

With the genome sequencing technology developing towards significant lower cost, shorter time and higher throughput, genomic data presents the explosive growth in recent years. Required by journals and funding agencies, sequencing data must be submitted to public database for accessibility. Inclusive of NCBI/SRA, EBI/SRA and DDBJ/DRA, International Nucleotide Sequence Database Collaboration (INSDC) has the capacity to store and publish the genome data from all over the world, but internet network transferring speed can not afford the big data transferring of long distance and different areas. To maximally overcome the disadvantages, we develop Genome Sequence Archive (GSA; http://bigd.ac.cn/gsa) in China for archiving and sharing the genomic data as well as accepting the submissions from all over the world. Since China has a powerhouse in generating big biological data, GSA can benefit the users especially the China local users in archiving sequencing & meta data and even making them real-time available for the worldwide scientific communities. GSA is currently one of the database resources with the international standards as INSDC which locates in BIG Data Center (BIGD; http://bigd.big.ac.cn) of Beijing Institute of Genomics (BIG), Chinese Academy of Sciences (CAS).

102 Towards comprehensive coverage and annotation of proteome space

Ramona Britto and Uniprot Consortium

The UniProt Knowledgebase (UniProtKB) endeavours to provide the scientific community with the most comprehensive catalog possible of protein sequence and functional information. To achieve this, we have put in place procedures that gather data accurately and consistently from source databases followed by extensive annotation and cross-referencing to other resources. At the heart of each UniProtKB record is a protein sequence - typically derived from the translation of a protein-coding gene on a sequenced genome. As the cost of sequencing continues to fall, the number of organisms with completely sequenced and annotated genomes is growing at an unprecedented rate. At UniProt we provide comprehensive protein-centric views of such genomes through the Proteomes portal (http://www.uniprot.org/proteomes/). The majority of currently available proteomes (45,162 proteomes, UniProt release 2016_01) are based on the translation of completely sequenced genomes submitted to the EMBL/GenBank/DDBJ databases or the International Nucleotide Sequence Database Collaboration (INSDC). Submitted genomes sometimes lack gene model predictions or have problems that prevent the generation of a non-redundant protein set. In the past, this has included important model organisms, such as Rattus norvegicus (Rat) and Zea mays (Maize). In other cases previously sequenced genomes are reannotated by an expert community for example, Triticum aestivum (Wheat); and these datasets are not always available through the INSDC. UniProtKB overcomes these issues by generating these proteomes in collaboration with groups such as Ensembl and model organism databases (MODs). Over the last few years we have established complementary pipelines for import of protein sequences from three alternate sources: Ensembl (vertebrates), Ensembl Genomes (invertebrates) and more recently WormBase ParaSite (helminth genomes). Further analysis has revealed that extending this approach to other databases such as VectorBase, FlyBase and NCBI RefSeq would bring us closer to capturing the vast majority of sequenced genomes. Accurate identification and incorporation of new, publically available, annotated genomes is a complex task and requires evaluation of many factors. Genome size and coverage, contig and scaffold N50 measures, availability of species-specific transcript and protein sequences are all used to evaluate candidate genomes for inclusion to UniProtKB. For particularly complex cases (such as parasite genomes), core genes and single-copy gene analysis measures are taken into account to verify completeness. At present all non-INSDC submissions are assessed by a curator and included manually. As of release 2016_01 (January 2016) UniProtKB contains the proteomes of 135 organisms from alternate sources. In addition to the import of new proteomes, maintenance and updating of existing proteomes to reflect improvements in genome assemblies and genebuild procedures is vital to the sustained growth of the proteomes project. This complex task is central to the functioning of UniProtKB and is overseen by a team of curators and programmers.We will present recent developments in this area and discuss ongoing work within the proteomes database aimed at maintaining the high standard and comprehensiveness of proteome data.

103 Assembly and curation of a clinical drug repurposing library for therapeutic discovery

Jodi Hirschman, Steven Corsello, Josh Gould, Patrick McCarren, Bang Wong, Jenny Liu, Mariya Khan, Stephen Johnston, Anita Vrcic, Aravind Subramanian, Josh Bittker and Todd Golub

There is wide and growing interest in repurposing drugs for new therapeutic applications. Examples of "repurposed" compounds include the known teratogen thalidomide, which has been found to be effective in the treatment of multiple myeloma, and the antibiotic erythromycin for treatment of impaired gastric motility (Musto, et al., 2008; Altomare, et al, 1997). Repositioning drugs for new applications is compelling as it offers a less expensive and more rapid alternative to the development of new untested compounds, since the former have already passed at least the first phase of clinical trials. A systematic repurposing effort requires a definitive collection of clinical drugs and a read-out of drug activity that allows prediction of new indications. To address the first need, we have assembled a repurposing library of approximately 5000 compounds containing over 3000 clinical drugs. One of the major challenges in assembling this collection has involved locating and verifying chemical structures for the compounds. We searched available databases for clinically-tested drugs and governmental regulatory agencies for approved ingredient lists and identified 11,000 small molecule compounds with unique disclosed drug structures that have been approved or are in clinical development. Those available for purchase (~5000) were annotated for clinical development status, mechanism of action, and targets, using publicly available databases and published literature as sources. Targets were mapped to PANTHER protein categories and MoAs were manually mapped to a standardized vocabulary of top-level and detailed terms. We are creating a web application in HTML5 and javascript to allow users to search the contents of the library by drug name, chemical structure, external database links, highest clinical development status obtained, standardized mechanism of action, and gene targets. Experiments are currently underway to test compounds of the repurposing library for their transcriptional effects in 10 cell lines, as well as for viability effects across 575 cell lines. This library is also being used by the NIH LINCS program.

104 Transcriptomics big data curation, sharing and re-analysis challenges: example of GTEx data integration into Bgee

Anne Niknejad, Amina Echchiki, Angelique Escoriza, Julien Roux, Sebastien Moretti, Marc Robinson-Rechavi and Frederic

B. Bastian

High-throughput sequencing technologies currently allow to produce amounts of data so vast that the entire biomedical field is facing new challenges, often previously underestimated, such as: long term storage and sharing of large datasets; capture and provision of accurate metadata; big data re-analyses and integration in various contexts. Datasets from "big data" projects in the field of transcriptomics are increasingly available, such as the Genotype-Tissue Expression (GTEx) dataset. For the development of the gene expression database Bgee (http://bgee.org/), it is a necessity to be able to integrate such large datasets. The aim of the Bgee database is to provide a reference of normal gene expression in animals, comparable between species, currently comprising human and 16 other species. The GTEx data, composed of thousands of RNA-Seq libraries, collected from more than 500 human subjects and 50 tissues, was essential to be integrated. We will share here the lessons learnt from the integration of GTEx into Bgee, and the solutions adopted, related to: the curation and standardization of abundant metadata; the acquisition and storage of terabytes of raw data; the gene expression quantification steps, leveraging the recent improvements in RNA-Seq analysis software and HPC experience; the integration and comparison of these data with other datasets and other species. We have notably conducted a work of re-annotation, using the metadata provided under restricted access, that allowed us to refine the annotations of anatomical structures, and to clarify some imprecisions; this also led us to request the inclusion of new terms into the Uberon anatomical ontology. Subjects and samples were also carefully filtered, in order to capture a high quality dataset of "normal" (healthy) gene expression in human, that might be of interest to a large community of researchers. In order to then process this dataset, we had to implement the use of new tools such as Aspera to acquire the data, or Kallisto to map the RNA-seq reads in a scalable way. The results were then standardized and reduced as qualitative patterns of gene expression, over anatomy, development and aging, to be integrated into Bgee. This approach allows to produce a biologically meaningful summary of such a huge amount of data; and to scale for integrating future "big data" projects to come, in the area of transcriptomics.Bgee is available at http://bgee.org/. Our work of re-annotation of GTEx is available at https://github.com/BgeeDB/expression-annotations.

105 UbiGRID: a resource for interactions and post-translational modifications in the ubiquitin-proteasome system

Rose Oughtred, Bobby-Joe Breitkreutz, Lorrie Boucher, Christie Chang, Jennifer Rust, Nadine Kolas, Lara O'Donnell, Chris Stark, Kara Dolinski, Mike Tyers and Andrew Chatr-Aryamontri

The covalent attachment of ubiquitin to substrate proteins controls the stability, interactions, activity and/or localization of much of the proteome. The canonical ubiquitination cascade proceeds by a three-step process: activation of ubiquitin as a thioester by an E1 enzyme, transfer to an E2 enzyme as a thioester intermediate, and conjugation to a lysine residue (or N-terminal amino group) on the substrate or the extending polyubiquitin chain by an E3 enzyme. Conversely, the extent of substrate ubiquitination is dynamically controlled by a host of deubiquitinating enzymes, which often act in balanced concert with E3 enzymes. The fate of the ubiquitinated substrate is determined by interactions with a host of ubiquitin binding domains, which can direct the substrate for degradation by the 26S proteasome or alter substrate localization, interactions or activity. The broad effects of the UPS on the proteome, and its connections to many disease states, have catapulted the UPS to the forefront of drug discovery in the pharmaceutical and biotechnology sectors. The therapeutic potential of the UPS has been largely underexplored due to insufficient understanding of the interdependence and redundancy between the different system components, the incomplete mapping of substrate-UPS system interactions and the largely unknown druggability of UPS enzymes. The goal of the UbiGRID curation project is to help fill this gap by comprehensively annotating the genetic and protein interactions of all UPS genes/proteins in humans, budding yeast and other model species. UbiGRID will serve as a centralized resource for three types of data: (i) an annotated reference set of UPS components organized into functional classes; (ii) comprehensive curation of genetic and protein interactions for all UPS genes; (iii) the annotation of ubiquitinated residues derived from mass spectrometry datasets.We have developed the most complete annotation of the core UPS machinery for human cells reported to date, which encompasses 1275 known and inferred system components. We recently completed the curation of 84,595 human protein interactions for these 1275 genes, as derived from 10,464 publications. Correspondingly, 31,886 yeast UPS protein interactions have been derived from 2,408 publications and 39,285 yeast genetic interactions from 2,420 publications. We have also captured documented 87,018 sites of ubiquitin modification for human proteins and 13,450 sites for yeast proteins. Collectively, this UPS interaction dataset should facilitate fundamental and applied discoveries in the UPS.

106 Towards 1000 Fungal Genomes: from data curation to high-throughput analysis

Igor Grigoriev

Future energy demands and environmental challenges can be addressed by learning from biological processes encoded in living organisms and microbial communities. Fungi are among the most powerful plant pathogens and symbionts as well as biomass decomposers. The Fungal Genomics Program of the US Department of Energy (DOE) Joint Genome Institute (JGI) is partnering with international scientific community to explore the fungal diversity in several large scale genomics initiatives.One of such initiatives, the 1000 Fungal Genomes project, is aimed at exploring diversity across the Fungal Tree of Life in order to understand fungal evolution, to build parts lists of genes, enzymes and pathways for biotechnological applications, and to provide references for environmental metagenomics. Its scale offers new challenges in data production, integration, analysis and requires a unique balance between automated high throughput analysis and manual curation techniques.This balance enables efficient integration of genomic and other omics data in the JGI fungal genomics resource MycoCosm (jgi.doe.gov/fungi), which currently contains over 600 fungal genomes and provides tools for comparative genomics and community-driven data curation.

107 The Disease Portals, Disease-Gene Annotation and the RGD Disease Ontology at the Rat Genome Database

G. Thomas Hayman, Stanley J. F. Laulederkind, Jennifer R. Smith, Shur-Jen Wang, Victoria Petri, Rajni Nigam, Marek Tutaj, Jeff De Pons, Melinda R. Dwinell and Mary Shimoyama

The Rat Genome Database (RGD; http://rgd.mcw.edu/) provides critical datasets and software tools to a diverse community of rat and non-rat researchers worldwide. To meet the needs of the many users whose research is disease oriented, RGD has created a series of Disease Portals and has prioritized its curation efforts on the datasets important to understanding the mechanisms of various diseases. Gene-disease relationships for three species, rat, human and mouse, are annotated to capture biomarkers, genetic associations, molecular mechanisms, and therapeutic targets. To generate gene-disease annotations more effectively and in greater detail, RGD initially adopted the MEDIC disease vocabulary from the Comparative Toxicogenomics Database and adapted it for use by expanding this framework with the addition of over 900 terms to create the RGD Disease Ontology (RDO). The RDO provides the foundation for, at present, ten comprehensive disease area-related dataset and analysis platforms at RGD, the Disease Portals. Two major disease areas are the focus of data acquisition and curation efforts each year, leading to the release of the related Disease Portals. Collaborative efforts to realize a more robust disease ontology are underway.

108 Atlas of Cancer Signaling Network interactive pathway database for data visualization and mathematical modeling in cancer biology

Inna Kuperstein, Eric Bonnet, Christophe Russo, Hien-Anh Nguyen, David Cohen, Laurence Calzone, Maria Kondratova, Eric Viara, Marie Dutreix, Emmanuel Barillot and Andrei Zinovyev

Studying reciprocal regulations between cancer-related pathways is essential for understanding signaling rewiring during cancer evolution and in response to treatments. With this aim we have constructed the Atlas of Cancer Signaling Network (ACSN), a resource of cancer signaling maps and tools with interactive web-based environment for navigation, curation and data visualization. The content of ACSN is represented as a seamless 'geographic-like' map browsable using the Google Maps engine and semantic zooming. The associated blog provides a forum for commenting and curating the ACSN maps content. The atlas contains multiple crosstalk and regulatory circuits between molecular processes implicated in cancer (Kuperstein et al, 2015). The integrated NaviCell web-based tool box allows to import and visualize heterogeneous omics data on top of the ACSN maps and to perform functional analysis of the maps. NaviCell web-based tool box is also suitable for computing aggregated values for sample groups and protein families and mapping this data onto the maps. The tool contains standard heatmaps, barplots and glyphs as well as the novel map staining technique for grasping large-scale trends in numerical values projected onto a pathway map. The NaviCell web service provides a server mode, which allows automating visualization tasks and retrieve data from maps via RESTfull (standard HTTP) calls. There is also a possibility of bindings to several programming languages as Python, R, Java (Bonnet et al, 2015). To demonstrate applications of ACSN and NaviCell we show a study on drug sensitivity prediction using the networks. We performed a structural analysis of Cell Cycle and DNA repair signaling network together with omics data from ovary cancer patients resistant to genotoxic treatment. Following this study we retrieved synthetic lethal gene sets and suggested intervention gene combinations to restore sensitivity to the treatment. In additional study we show how epithelial to mesenchymal transition (EMT) signaling network from the ACSN collection has been used for finding metastasis inducers in colon cancer through network analysis. We performed structural analysis of EMT signaling network that allowed highlighting the network organization principles and complexity reduction up to core regulatory routs. Using the reduced network we modeled single and double mutants for achieving the metastasis phenotype. We predicted that a combination of p53 knock-out and overexpression of Notch would induce metastasis and suggested the molecular mechanism. This prediction lead to generation of colon cancer mice model with metastases in distant organs. We confirmed in invasive human colon cancer samples the modulation of Notch and p53 gene expression in similar manner as in the mice model, supporting a synergy between these genes to permit metastasis induction in colon (Chanrion et al, 2014).Kuperstein I et al, Atlas of Cancer signaling Network: navigating cancer biology with Google Maps Oncogenesis. doi: 10.1016/j.bbrc.2015.06.094 (2015) Bonnet E et al, NaviCell Web Service for network-based data visualization. Nucleic Acids Res. doi:10.1093/nar/gkv450 (2015) Chanrion et al, Notch activation and p53 deletion induce EMT-like processes and metastasis in a novel mouse model of intestinal cancer. Nature Communications 5:5005. doi: 10.1038/ncomms6005 (2014)

109 HPIDB 2.0: a curated database for host-pathogen interactions

Mais Ammari, Cathy Gresham, Fiona McCathy and Bindu Nanduri

Identification and analysis of host-pathogen interactions (HPI) data has a huge impact on disease treatment, management and prevention. HPIDB 2.0 (http://www.agbase.msstate.edu/hpi/main.html) provides a unified query interface for HPI information, and contains 43,276 manually curated entries in the current release. Since the first HPIDB release in 2010, multiple enhancements to HPIDB data and interface services were made to facilitate both the identification and functional analysis of HPI data. Notably, HPIDB 2.0 now provides targeted biocuration of HPI data. Annotations provided by HPIDB curators meet International Molecular Exchange consortium standards to provide detailed contextual experimental information and facilitate data sharing. In addition to curation, HPIDB 2.0 integrates HPI from existing external sources and contains tools to infer additional HPI where annotated data is scarce. Our data collection approach ensures HPIDB 2.0 users access comprehensive HPI data from a wide range of pathogens and their hosts (567 pathogen and 68 host species, as of December 2015) and avoid the time-consuming series of steps required to integrate, standardize, and annotate HPI data. The data updates are accompanied with enhanced web interface that allows the users to search, visualize, analyze and download HPI data. Perhaps most noticeably for our users, we have expanded the HPIDB 2.0 results to display additional interaction information, associated host and/or pathogen Gene Ontology functions and network visualization. All HPIDB 2.0 data are updated regularly, are publicly available for direct download, and are disseminated to other MI resources. Our future goals for HPIDB include broadening the number of pathogens for which experimentally derived manual curation HPI data is available and enabling the end user to evaluate the quality of transferred homologous HPI for improved computational HPI prediction.

110 Human protein variants in UniProtKB/Swiss-Prot: improving access to knowledge through standardized annotations.

Maria Livia Famiglietti, Lionel Breuza, Teresa Neto, Sebastien Gehant, Nicole Redaschi, Sylvain Poux, Lydie Bougueleret, Ioannis Xenarios and Uniprot Consortium

UniProtKB/Swiss-Prot (http://www.uniprot.org) provides the scientific community with a collection of information, expertly curated from the scientific literature, on protein variants. Priority is given to single amino-acid polymorphisms (SAPs) found in human proteins, their functional consequences and association with diseases. UniProt release 2016_01 includes over 74,000 human SAPs, 38,515 of which are enriched by annotations in free-text describing involvement in disease and functional characteristics of the variant. To ease access to this knowledge and to make it computer readable, we are restructuring these annotations using controlled vocabulary. By combining terms from Variation Ontology (VariO) and Gene Ontology (GO), we can describe the large spectrum of effects caused by SAPs on proteins. We use VariO terms to indicate which protein property is affected, such as its structure, expression, processing, and function. GO terms are used to specify which biological process, protein function, or subcellular location are impacted. A limited number of controlled attributes complete the annotations, defining how the protein property is affected, e.g. increased, decreased or missing. Currently, most SAPs with at least 1 annotation have been reviewed, producing close to 7,000 structured functional annotations. We plan to provide this new structured format to our users as soon as possible.

111 Ontological Reasoning for Immunological Diseases of the IEDB using the DO

Randi Vita, Elvira Mitraka, James A. Overton, Lynn M. Schriml and Bjoern Peters

The IEDB has been describing immune epitope experiments, as presented in the scientific literature, for more than 10 years and has accumulated a significant dataset, representing what is currently known in the field of immune epitopes. The goal of this project was to accurately model the disease states presented in this literature in a logical and consistent manner. To achieve this goal, we reviewed all disease states described in our literature set, determining the method of exposure (e.g. natural infection), type of disease (e.g. allergy), site of disease (e.g. respiratory tract) and the immunogen (e.g. Plasmodium falciparum) in order to establish logical rules to define disease states. This work generated clear logical definitions for disease states and enabled enforceable validation rules, as well as identifying errors within our dataset. In order to share the results of this work with overlapping domains and to make the IEDB more interoperable with other resources, the resulting disease states and their definitions were submitted to the Human Disease Ontology (DO). DO is the well-established standardized ontology for human disease. Through collaboration between the IEDB and DO, these diseases were logically modeled within the DO. Thus, the IEDB can now incorporate a subset of DO as an OWL file to create a searchable disease tree for our curators and end users to easily view disease information in a hierarchical tree format. Additionally, the reasoned ontology provides validation rules that are being incorporated into the IEDB's curation interface to improve accuracy of curated data. Going forward, as new diseases are encountered in the literature, the process will be repeated. Our hope is that other resources will increasingly utilize these same diseases within DO and the richness of our data will grow, identifying overlapping datasets and allowing new scientific conclusions.

112 Curating Cancer Gene Census

Sari Ward, Zbyslaw Sondka, Sally Bamford, Charlotte G. Cole, David M. Beare, Nidhi Bindal, Tisham De, Simon A. Forbes, John Gamble, Mingming Jia, Chai Yin Kok, Kenric Leung and Peter J. Campbell

The Cancer Gene Census is an ongoing effort to catalogue those genes for which somatic mutations have been causally implicated in cancer. Originally published in 2004 (Futreal et al, 2004), the Census has been continued and maintained by COSMIC, the Catalogue Of Somatic Mutations In Cancer, which is the world's largest and most comprehensive resource to explore the impact of somatic mutations in human cancer. Currently, 3% of all human genes are implicated in cancer development via somatic mutations or gene fusions. Out of 571 Census genes, missense mutations have been reported in 539, nonsense mutations in 366 and inactivating frameshift mutations in 331 genes. Various other types of mutations were found in 318 Census genes, and 353 genes were involved in gene fusions. For a gene to be included in the Census, at least two independent publications need to exist showing mutations in this gene in primary patient material, and these should have evidence of the somatic origin of at least a subset of mutations based on analysis of normal tissue from the same individuals. As germline fusions are relatively uncommon, cancer genes involved in fusions may be included without definite evidence of somatic origin. Also, single reports of novel fusions in rare tumours are included. Further inclusion and exclusion rules are applied when considering new genes for inclusion. The Census is updated continuously with new genes and related information including the tumour types in which mutations are found, classes of mutation that contribute to oncogenesis, molecular mechanism of the gene in cancer and other genetic properties. The Census is a manually curated summary of most relevant information on cancer driving genes and their somatic mutations gathered in COSMIC database and brings together the expertise of a dedicated curation team, in house and external cancer scientists and a wide user community. Since 2015 it has also become a part of the Centre for Therapeutic Target Validation (CTTV) platform (https://www.targetvalidation.org) which brings together information on the relationships between potential drug targets and diseases using evidence from multiple data types that can be relevant to target identification and prioritisation. The Census is available from the COSMIC website for online use or download at: (http://cancer.sanger.ac.uk/census).

113 Deriving Gene-to-Phenotype and Gene-to-Disease from Mouse Genotype Annotations: Challenges and Solutions

Susan Bello, Janan Eppig and The Mgi Software Group

Phenotypes and diseases are emergent properties of whole organisms. At Mouse Genome Informatics (MGI, www.informatics.jax.org), we curate models of human disease to the mice used in experiments, specifically using the key allele pairs and strain backgrounds that define the full genetic makeup of the mice. In order to derive gene-to-phenotype and gene-to-human disease model relationships from annotated mouse models, we need to identify models that contain mutations in only a single causative gene. These derived gene annotations can then be used to provide users with a high-level summary of gene function and be used in candidate disease gene analysis. Filtering the various kinds of genotypes to determine which phenotypes are caused by a mutation in a particular gene can be a laborious and time-consuming process. At MGI we have developed a gene annotation derivation algorithm that computes gene-to-phenotype and gene-to-disease annotations from our existing corpus of annotations to genotypes (allele pairs and strain background). This algorithm differentiates between simple genotypes with causative mutations in a single gene and more complex genotypes where mutations in multiple genes may contribute to the phenotype. The process identifies alleles functioning as tools (e.g., reporters, recombinases) and filters these out. Several improvements in allele descriptions in MGI have been used to refine the accuracy of this algorithm. These include 1) creation of allele attributes that are used to identify tools, 2) introduction of relationships between alleles and the genes expressed by the allele, 3) introduction of relationships between multi-genic alleles and all genes in the mutation region. Using this algorithm, derived gene-to-phenotype and gene-to-disease annotations were created for 16,000 and 2,100 mouse markers, respectively, starting from over 57,900 and 4,800 genotypes with at least one phenotype and disease annotation, respectively. Implementation of this algorithm provides consistent and accurate gene annotations across MGI and provides a vital time-savings relative to manual annotation by curators. Supported by NIH grant HG000330.

114 DMDD Project: Curation and visualisation of 3D image data for mutant phenotypes

Robert Wilson, Julia Rose, Stefan Geyer, Lukas Reissig, Dorota Szumska, Andrew Cook, Wolfgang Weninger, Cecilia Mazzeo, Jacqui White, Fabrice Prin and Tim Mohun

The Deciphering the Mechanisms of Developmental Disorders (DMDD) consortium is a research programme characterising mouse lines carrying a targeted mutation that show embryonic and perinatal lethality when the mutation is homozygous. One of the goals of the project is to identify lines useful to developmental biologists and clinicians as animal models for investigating the basis of human developmental disorders. Approximately a third of all mouse strains that carry a null mutation show homozygous recessive embryonic or perinatal lethality, and among this group at least 60% show structural defects in one or more organ system that can be identified in histological sections by conventional microscopy. The DMDD project studies embryos that survive to at least mid gestation using a combination of comprehensive high resolution episcopic microscopy (HREM) for 3D imaging, and tissue histology to identify abnormalities in developing organ and placental structures. The images we collect are screened systematically for morphological defects by a team of developmental biologists and anatomists. The mutant phenotypes observed are recorded by using terms from the Mammalian Phenotype Ontology, or our own controlled vocabulary that enables us to document phenotypes in a systematic fashion prior to representation of the phenotype in the ontology. For 3D image datasets we capture the location at which the phenotype was observed through a plugin we developed for the open source image processing and visualisation software package Osirix. This plugin allows curators to export, import and merge sets of ontology terms and comments associated with points in 3D space, which makes it possible to carry out phenotype annotation at several sites. The image data and the phenotypes we have scored are available through the project website (http://dmdd.org.uk). The search function of the website enables end users to navigate directly to the 3D location within the image data that is the basis of the curated phenotype statement, and the stackviewer interface that displays the image also allows this section to be compared to similar ones from other embryos.

115 Grouping transgenic construct alleles in FlyBase using a new 'class' vocabulary.

Gillian Millburn

FlyBase uses an 'allele class' controlled vocabulary to type classical mutant alleles according to their function, recording for example whether they are 'hypomorphic' or 'amorphic' loss of function alleles, or 'neomorphic' or 'antimorphic' gain of function alleles. Using this controlled vocabulary allows users to easily search for particular types of classical mutant alleles. However, we currently have no equivalent controlled vocabulary for alleles that represent transgenic constructs introduced into flies, with information describing the nature of the transgenic construct only being captured as free text. We describe our initial attempts at formulating a 'class' controlled vocabulary for transgenic construct alleles and assess how this fits onto our existing set of alleles. We also discuss how we hope to use this new controlled vocabulary to improve summarisation of phenotype and genetic interaction data on the FlyBase website. Finally we examine how we can use this new controlled vocabulary to computationally derive new types of information from our existing curation of phenotype and genetic interaction data, for example to derive functional complementation statements.

116 Harmonizing Disease Annotation of Mouse and Rat Models Through the Human Disease Ontology

Lynn Schriml, Mary Shimoyama, Elvira Mitraka, Susan Bello, Stanley Laulederkind, Cynthia Smith and Janan Eppig

The use of model organisms to study the mechanisms of human disease is growing rapidly. Concomitant with this growth is the need for a disease ontology to facilitate comparisons of research findings and disease profiles between human and model organisms and to aid in identifying the underlying genetic, genomic and physiological mechanisms of disease. The Mouse Genome Database (MGD, http://www.informatics.jax.org) and the Rat Genome Database (RGD, http://rgd.mcw.edu) are teaming up with the Disease Ontology (DO, http://www.disease-ontology.org) project to harmonize disease annotation through collaborative review of MGD (OMIM), RGD (MEDIC) and DO disease terms and to update and enhance the structure and content of the DO to improve its capacity to support cross-organism representation of disease. Our major goal is to foster the adoption of a shared, robust DO for MGD and RGD through the enhancement of DO to support MGD and RGD disease annotations. As an added benefit, DO will become more comprehensive and useful to other projects annotating various types of data generated from a wide variety of experimental and clinical investigations. The ability to consistently represent disease associations across data types from the cellular to the whole organism, generated from clinical and model organism studies will facilitate data integration, data mining and comparative analysis. The progressive enrichment of the DO and successful adoption of DO for disease annotation by MGD and RGD will demonstrate its potential use across organisms and will encourage other groups to look at DO as a standard for their disease annotation needs. In addition, use of DO will greatly increase the potential for interoperability between MGD and RGD systems at the disease annotation level and provide the human genetics/genomics community with a consistent way to query for rodent disease associations.Disease Ontology Database URL: http://www.disease-ontology.orgMouse Genome Database URL: http://www.informatics.jax.orgRat Genome Database URL: http://rgd.mcw.edu

117 Enhancing the Human Phenotype Ontology for use by the Layperson

Nicole Vasilevsky, Mark Engelstad, Erin Foster, Sebastian Kohler, Chris Mungall, Peter Robinson and Melissa Haendel

Many diseases present with distinct phenotypes, making descriptions of phenotypes valuable for identifying and diagnosing human diseases. The Human Phenotype Ontology (HPO) was developed to provide a structured vocabulary containing textual and logical descriptions of human phenotypes. The HPO is used for phenotype-genotype alignment in systems like the Monarch Initiative to provide disorder prediction, variant prioritization, and patient matching between known diseases and model organisms. Here we describe recent work to extend the utility of the HPO through the systematic addition of approximately 6,000 synonyms. Until now, most of the HPO synonyms were composed of clinical terms unfamiliar to patients. For example, a patient may know they are 'color-blind', but may not be familiar with its official phenotype term 'Dyschromatopsia'. Therefore, our goals is to add synonyms in "layperson-ese" so that HPO can be used by patients as well as basic research scientists and clinicians to help improve disease characterization and diagnosis. We systematically reviewed current HPO classes (approximately 12,000) and assigned layperson synonyms to each class where applicable. The layperson synonyms refer to colloquial terms used to describe phenotypic features associated with medical conditions. Each layperson synonym was annotated to indicate its special status, then classified as either exact (precise); broad (more general); narrow (more specific); or related (associated). The review process included various methods of identifying and validating possible layperson synonyms. We first queried the HPO to avoid duplicate terms. We then batched similar kinds of terms together, such as those related to bone abnormalities, to maintain consistent synonym terminology. For example, the phenotypes of the femur were assigned layperson synonym 'of thigh bone' and morphological abnormalities were described as 'abnormal shape of '. We consulted online resources (e.g., Wikipedia, Mayo Clinic) as well as specialized resources (e.g., Uberon, Gene Ontology) to find additional synonyms. As a quality control measure, we reviewed each other's work, consulted with clinical experts when necessary, and queried Google for the assigned layperson term to verify that it retrieved the appropriate medical term and was in use. Some challenges of assigning layperson synonyms involved reconciling lay terms with the logic and structure of the HPO and determining the best mechanism to validate the lay synonyms. Additionally, not every term has a lay synonym or it may already exist in the HPO, such as 'widow's peak' or 'hitch hiker's thumb'. Finally, some terms have complicated medical terminology, like 'short distal phalanx of first finger', for which a single layperson term is difficult to establish without using the definition of the term. The addition of layperson synonyms increases the usability of the HPO, making it useful for data interoperability across clinicians and patients. Additionally, this work will enable crowdsourcing by citizen scientists. The layperson synonyms will be available as a modular import file for the HPO and are due to be released in the Spring of 2016.

118 UniProt at EMBL-EBI's role in CTTV: contributing to improved disease knowledge

Barbara Palka, Daniel Gonzalez, Edd Turner, Xavier Watkins, Maria Martin and Claire O'Donovan

At its core, UniProt (the Universal Protein Resource) is a collection of protein sequences and protein-related information. One of the main components of UniProt is the UniProt Knowledgebase. The goal of UniProtKB is to organise and annotate information about protein function and sequence, providing a comprehensive overview of available information. Annotation efforts are both automatic and manual. Information is computationally added to uncharacterised sequences in the unreviewed TrEMBL section of UniProtKB while experimentally characterised proteins undergo a process of manual expert curation before entering the reviewed Swiss-Prot section. Manual curation consists of a critical review of experimental and predicted data for each protein as well as manual verification of each protein sequence. The range of captured information covers protein function, interactions, catalytic activity, patterns of expression and disease associations, just to name few. All curation efforts are presented to the scientific community in the form of a comprehensive, high-quality and freely accessible resource.Launched to the public in December 2015, The Centre for Therapeutic Target Validation (CTTV) platform is a pioneering public-private research initiative between GlaxoSmithKline, EMBL-EBI and the Wellcome Trust Sanger Institute. CTTV aims to make use of recent progress in genome sequencing and the potential of "big data" to improve the success rate for new drug discovery. UniProt is an integral part of this platform and supplies valuable information about target function to the protein profile section of the website. It also provides the graphical overview of protein features in the form of an interactive viewer.Data integration across multiple resources contributing to CTTV content and the creation of the map of complex relationships between diseases is possible as a result of mapping diseases to the terms in the Experimental Factor Ontology (EFO). UniProt curators have contributed to the creation of the disease associations by manually mapping over a thousand rare diseases that could not be mapped automatically. The mapping is an ongoing process and UniProt curators continue to map novel diseases in UniProt to common disease phenotype terms in the EFO. CTTV has many current areas of interest ranging from cancer to auto-immune diseases. UniProt curators are currently involved in updating target entries for proteins associated with two inflammatory bowel disease (IBD) conditions Crohn's disease and ulcerative colitis.Having a collection of the most up-to-date, cutting-edge experimental data presented in a well-organised target- or disease-centric manner will contribute to the effort to decipher disease-causing factors and help in the search for new treatments.

119 Annotation of functional impact of mutations in cancer predisposing genes

Isabelle Cusin, Monique Zahn, Valerie Hinard, Pierre Hutter, Amos Bairoch and Pascale Gaudet

Characterization of the phenotypic effect of mutations provides evidence on which variants of unknown significance

(VUS) can be more precisely evaluated. In this context we have annotated the phenotypes caused by mutations in BRCA1, BRCA2, PALB2, MLH1, MSH2, MSH6, PMS2, APC, MUTYH associated with increase susceptibility to the most prevalent hereditary breast and colorectal cancers.Using the information derived from publications, the functional impact of variants was captured, resulting in thousands of different annotations. Annotations are organized in triplets, consistent with the RDF model. The annotations are composed of terms from ontologies and controlled vocabularies to ensure consistency in descriptions and support computational analysis. Each annotation is supported by detailed experimental evidence.Well characterized and assessed functions of the target proteins are captured. For instance for BRCA1 and BRCA2 this includes their ubiquitin-protein ligase and transcriptional regulation activities, their role in DNA repair in response to DNA damage. Important in term of phenotypic outcome are direct interactions with UBE2D1, BARD1 and BRIP1. For genes involved in DNA mismatch repair, this includes their overall mismatch repair ability, mismatch complex formation, as well as the individual steps of mismatch binding, ATPase activity, ADP/ATP exchange, etc. These annotations provide a comprehensive resource on the phenotypic outcome of cancer predisposing genes in a level of detail and structure superior to what is found in other databases.