25

PhenoDigm: analyzing curated annotations to associate animal models with human diseases

Damian Smedley[1], Anika Oellrich[1], Sebastian Kohler[2], Barbara Ruef[3], Sanger Mouse Genetics Project[1], Monte Westerfield[3], Peter Robinson[2], Suzanna Lewis[4]

[1] Wellcome Trust Sanger Institute, United Kingdom
[2] Institute for Medical Genetics and Human Genetics, United Kingdom
[3] Department of Biology, United States of America
[4] Lawrence Berkeley National Laboratory, United States of America

Presenter: Damian Smedley

Model organisms are studied with the ultimate goal of translating insights into valuable knowledge about human disease to facilitate treatment and early screening for diseases. Recent advancement in technologies allow for rapid generation of models with a targeted range of genotypes as well as their characterisation through high throughput phenotyping. As an abundance of data become available, only systematic analysis allows valid conclusions to be drawn from these data and transferred to human diseases. Due to the volume of data, automated methods are preferable, allowing for a reliable analysis of the data and providing evidence about possible gene--disease associations. Here, we propose such a method, called PhenoDigm (Phenotype comparisons for DIsease Genes and Models), that provides evidence about gene--disease associations by analysing phenotype information. PhenoDigm integrates data from a variety of model organisms and, at the same time, employs several intermediate scoring methods to identify only strongly data-supported gene candidates for human genetic diseases. We show results of an automated evaluation as well as selected manually assessed examples that support the validity of PhenoDigm. Furthermore, we provide a guidance on how to browse the data with PhenoDigm's web interface and illustrate its usefulness in guiding research. Database URL:
http://www.sanger.ac.uk/resources/databases/phenodigm/

26

The new modern era of yeast genomics: Community sequencing and the resulting annotation of multiple Saccharomyces cerevisiae strains at SGD

Stacia Engel, J. Michael Cherry

Stanford University, United States of America

Presenter: Stacia Engel

The first completed eukaryotic genome sequence was that of the yeast Saccharomyces cerevisiae, and the Saccharomyces Genome Database (SGD; http://www.yeastgenome.org/) is the original model organism database. SGD remains the authoritative community resource for the S. cerevisiae reference genome sequence and its annotation, and continues to provide comprehensive biological information correlated with S. cerevisiae genes and their products. A diverse set of yeast strains has been sequenced to explore commercial and laboratory applications, and a brief history of those strains is provided. The publication of these new genomes has motivated the creation of new tools, and SGD will annotate and provide comparative analyses of these sequences, correlating changes with variations in strain phenotypes and protein function. We are entering a new era at SGD, as we incorporate these new sequences and make them accessible to the scientific community, all in an effort to continue in our mission of educating researchers and facilitating discovery.

27

PhenoMiner: Curating Annotations with Five Ontologies/Vocabularies Simultaneously

Stanley Laulederkind

Medical College of Wisconsin, United States of America

Presenter: Stanley Laulederkind

The Rat Genome Database (RGD) is the premier repository of rat genomic and genetic data and currently houses over 40,000 rat gene records as well as human and mouse orthologs, over 2000 rat and 1900 human quantitative trait loci (QTLs) and more than 2900 rat strains. Biological information curated for these data objects includes disease associations, phenotypes, pathways, molecular functions, biological processes and cellular components. Recently, a project was initiated at RGD to incorporate quantitative phenotype data for rat strains, in addition to the currently existing qualitative phenotype data for rat strains, QTLs, and genes. A specialized curation tool was designed to generate manual annotations with up to six different ontologies/vocabularies to describe a single experimental value from the literature. Concurrently, three of those ontologies needed extensive development to move the curation forward. The curation interface development, as well as ontology development, was an ongoing process during the early stages of the PhenoMiner curation project.

28

ProtocolNavigator: repairing the data curation and consumption imbalance

Imtiaz Khan[1], Adam Fraser[2], Mark-Anthony Bray[2], Paul Smith[1], Anne Carpenter[2], Rachel Errington[1]

[1] Cardiff University, United Kingdom
[2] Broad Institute of Harvard and MIT, United States of America

Presenter: Imtiaz Khan

Despite the exponential growth of biocuration infrastructures and databases, consumption or reuse of biological data beyond the proximity of the data originator remains rare. This imbalance is primarily due to the lack of effective communications among researchers. Markup and other structured languages utilized by current biocuration approaches enable researchers to share information; but these do not resonate with the day-to-day data curation culture, nor do they ease information interpretation for researchers who wish to reuse or assess the data. This is particularly prominent in an interdisciplinary research context, where the variability of methodologies, curation culture, and terminologies are highly idiosyncratic. Addressing this reality, we have developed ProtocolNavigator – a virtual laboratory environment that allows emulation of real life laboratory actions as the basis for curation. The emulation leads to the automatic representation of a time-integrated, interactive map of an experiment that includes action patterns, manipulations, and data acquisition signified by symbols. Association and sequential analysis of these symbols divulge patterns, which in turn provide a language-independent visual perception of experimental design. Navigation through this design reveals provenance trails for data and metadata; importantly, this interaction delivers contextualization, which facilitates knowledge abstraction and assessment. The fully navigable map can be shared with colleagues for design modification and optimization; also it can be converted and printed into a text format for publication. Our initial survey indicates that this mapping format facilitates intelligible data consumption and knowledge sharing without the use of language. ProtocolNavigator has enabled us to curate and compare practice variation – a quantitative approach for establishing best practice. This we believe a crucial social factor for wider community engagement and uptake.

29

Cataloging the biomedical world of pain through semi-automated curation of molecular interactions

Daniel Jamieson[1], Phoebe Roberts[2], David Robertson[1], Ben Sidders[3], Goran Nenadic[1]

[1] University of Manchester, United Kingdom
[2] Computational Sciences Center of Emphasis, Pfizer, United Kingdom
[3] Neusentis, UK, United Kingdom

Presenter: Daniel Jamieson

The vast collection of biomedical literature and its continued expansion has presented a number of challenges to researchers who require structured findings to stay abreast of and analyze molecular mechanisms relevant to their domain of interest. By structuring literature content into topic-specific, machine-readable databases, the aggregate data from multiple articles can be used to infer trends that can be compared and contrasted to similar findings from topic-independent resources. Our study presents a generalized procedure for semi-automatically creating a custom topic-specific molecular interaction database through the use of text mining to assist manual curation. We apply the procedure to capture molecular events that underlie 'pain', a complex phenomenon with a large societal burden and unmet medical need. We describe how existing text mining solutions are used to build a pain-specific corpus, extract molecular events from it, add context to the extracted events, and assess their relevance. The pain-specific corpus contains 765,692 documents from Medline and PubMed Central, from which there are 356,499 unique, normalized molecular event chains, with 261,438 single events and 93,271 molecular interactions supplied by BioContext. Event chains are annotated with additional contexts which collectively provide detailed insight into how that event chain is associated with pain. The extracted relations are visualized in a wiki platform (wiki-pain.org) that enables more efficient manual curation and exploration of the molecular mechanisms that underlie pain. Curation of 1,500 grouped event chains ranked by pain relevance revealed 613 accurately extracted unique molecular interactions that in the future can be used to study the underlying mechanisms involved in pain. Our approach demonstrates that combining existing text mining tools with domain-specific terms and wiki-based visualization can facilitate rapid curation of interactions to create a custom database.

30

A database for curating the associations between killer-cell immunoglobulin-like receptors and diseases in worldwide populations

Louise Takeshita[1], Faviel Gonzalez-Galarza[1], Eduardo Santos[2], Maria Helena Maia[2], Mushome Rahman[1], Syed Zain[1], Derek Middleton[3], Andrew Jones[1]

[1] Institute of Integrative Biology, University of Liverpool, United Kingdom
[2] Human and Medical Genetics, Federal University of Pará, Brazil
[3] Transplantation Immunology, Royal Liverpool and Broadgreen University Trust and University of Liverpool, United Kingdom

Presenter: Louise Takeshita

The killer cell-immunoglobulin-like receptors (KIR) play a fundamental role in the innate immune system, through their interactions with human leukocyte antigen (HLA) molecules, leading to the modulation of activity in natural killer (NK) cells, mainly related to killing pathogen infected cells. KIR genes are hugely polymorphic both in the number of genes an individual carries and in the number of alleles identified. We have previously developed the Allele Frequency Net Database (AFND, http://www.allelefrequencies.net), which captures worldwide frequencies of alleles, genes and haplotypes for several immune genes, including KIR genes, in healthy populations, covering over four million individuals. Here, we report the creation of a new database within AFND, named KIR and Diseases Database (KDDB), capturing a large quantity of data derived from publications in which KIR genes, alleles, genotypes and/or haplotypes have been associated with infectious diseases (e.g. hepatitis C, HIV, malaria), autoimmune disorders (e.g. type I diabetes, rheumatoid arthritis), cancer and pregnancy-related complications. KDDB has been created through an extensive manual curation effort, extracting data on more than a thousand KIR-disease records, comprising more than 50,000 individuals. KDDB thus provides a new community resource for understanding not only how KIR genes are associated with disease, but also, by working in tandem with the large data sets already present in AFND, where particular genes, genotypes or haplotypes are present in worldwide populations or different ethnic groups. We anticipate that KDDB will be an important resource for researchers working in immunogenetics.

31

The Protein Model Portal - a comprehensive resource for protein structure information

Juergen Haas, Konstantin Arnold, Florian Kiefer, Lorenza Bordoli, Torsten Schwede

Swiss Institute of Bioinformatics/Biozentrum University of Basel, Switzerland

Presenter: Juergen Haas

The Protein Model Portal (PMP) has been developed to foster effective use of molecular models in biomedical research by providing convenient and comprehensive access to structural information for a specific protein. Both experimental structures and theoretical models for a given protein can be searched simultaneously and analyzed for structural variation. By providing a comprehensive view on structural information, PMP offers the opportunity to apply consistent assessment and validation criteria to the complete set of structural models available for a specific protein. PMP is an open project to ensure that new methods developed by the community can be made available, for example new modeling servers and model quality estimation programs for model validation. The accuracy of the participating servers is continuously evaluated by the CAMEO (Continuous Automated Model EvaluatiOn) system. The Protein Model Portal thus offers a unique interface to visualize structural coverage of a protein by both theoretical models and experimental structures, allowing straightforward assessment of the model quality and hence their utility. The portal is updated regularly and actively developed to include latest methods in the field of computational structural biology. Visit us at www.proteinmodelportal.org!

33

TermGenie - Granting Biocurators' Wishes for the Gene Ontology

Heiko Dietze[1], Tanya Berardini[2], Rebecca Fougler[3], David Hill[4], Jane Lomax[3], Paola Roncaglia[3], Chris Mungall[1]

[1] Genomics Division, Lawrence Berkeley National Laboratory, United States of America
[2] The Arabidopsis Information Resource, United States of America
[3] European Bioinformatics Institute, United Kingdom
[4] Mouse Genome Informatics, The Jackson Laboratory, United States of America

Presenter: Heiko Dietze

A common bottleneck for biocuration is the generation of new ontology classes (terms). These classes are used to describe detailed aspects of a domain or to enhance a domain that is not well-represented in the ontology. The creation of a new class is not a trivial task. It must be ensured that the class does not already exist, that there are appropriate relations with respect to the overall structure of the ontology, and that there are textual definitions, supporting references, and synonyms. Here we present TermGenie, a web-based tool, which uses a pattern-based approach and reasoning to automate the creation of new ontology class. Using a TermGenie template, curators can verify and create a new class and permanent identifier quickly. The new class will have the appropriate meta data. Furthermore, the placement of the new class and the update of existing relations is done using the OWL reasoner ELK. Due to this pattern-based approach, the final review of new terms by senior editors is straightforward. The GeneOntology (GO) currently provides a TermGenie installation with 16 templates. Using templates, annotators and editors have created more than 3100 new GO classes in 2 years. Biocurators frequently need to create new GO classes that use external ontologies, such as ChEBI (Chemical Entities of Biological Interest). There are currently 9 templates using ChEBI, including those for metabolism, binding, transport and response to a chemical stimulus. For the integration in biocuration tools, there is a web service to verify the state of requested classes. This service is already integrated into the EBI annotation tool, Protein2GO, allowing curators to use newly created classes immediately for annotation. TermGenie is open source and is available here: http://termgenie.org The GO TermGenie server is available here: http://go.termgenie.org

34

Biocuration of next-generation sequence data: Infrastructure and proof-of-principle using cancer genomics data.

Raja Mazumder[1], Charles Cole[1], Vahan Simonyan[2]

[1] George Washington University, United States of America
[2] Food and Drug Administration, United States of America

Presenter: Raja Mazumder

Sequencing technologies have resulted in petabytes of scattered data, decentralized in archives, databases and sometimes even in isolated hard-disks that is inaccessible for browsing and analysis. Additionally, there is a lack of curated information in NGS data repositories such as NCBI SRA and CGHub. In the future, it is expected that curated secondary databases will help organize some of the Big Data, similar to what RefSeq and UniProtKB/Swiss-Prot have done for GenBank, with additional higher level databases such as Pfam and KEGG grouping objects into functional groups, biological networks and processes. To develop curated secondary databases from primary NGS data one needs low-cost infrastructure, algorithms and validated workflows that biocurators and community annotators can use. To address the above challenges, we have implemented novel NGS data compression, storage and processing infrastructure. Our approach includes a low-cost High-performance Integrated Virtual Environment (HIVE) private cloud at GWU and US FDA that provides a secure architecture of virtualized services integrated with highly parallelized computational algorithms. HIVE's ultra-fast mapping algorithm allows mapping of reads to reference genomes obtained using NCBI E-utils which then can be analyzed to calculate coverage, SNPs and a variety of other features. To assist the curation process we implemented a curation interface that has integrated annotations from UniProtKB-Swiss-Prot, RefSeq/CDD, Pfam, PANTHER and KEGG. Such integration allows curators to easily identify the effects of variation on active site, binding site, domains and pathways. As a proof-of-principle we have curated 12 datasets from NCI cancer genomics program. [Three publications that describe different aspects of this work: PMID: 22586465; http://dx.doi.org/10.1016/j.gpb.2012.10.003 ; Proteome-wide analysis of nsSNVs in active sites of human proteins, FJ-12-0816.R1, FEBS J in print]

41

UniProt databases tryptic search space

Emanuele Alpi, Johannes Griss, Alan Wilter Sousa da Silva, Benoit Bely, Daniel Ríos, Hermann Zellner, Rui Wang, Juan Antonio Vizcaino, Maria-Jesus Martin

EMBL-EBI, United Kingdom

Presenter: Emanuele Alpi

Most of the current MS driven bottom-up proteomics workflows exploit collections of sequences to match peptide sequences to experimental spectra and infer proteins to which peptides belong to. Trypsin is the most used cleaving agent in these types of workflows. A clear view of the tryptic search space of UniProt DBs can help scientists to select the most suitable collection. Comparing tryptic search spaces from different DBs assists in pinpointing differences between them and understanding the reasons behind. Integrating this information with the peptide-level identifications coming from MS-based proteomics repositories is a way of adding annotations to the corresponding proteins, so this flow of information can go in both directions. Tryptic search space of UniProt "complete proteome" sets ("canonical + isoforms" and "only canonical") for twelve model organisms available from UniProt ftp were compared with the corresponding IPI, RefSeq and Ensembl DBs. Natural variants for human and mouse were added in the comparisons and extended to whole UniProt sequence content and to the corresponding UniRef100 organism-specific DBs. Variation information was taken into account as a whole and as a subset of only the disease-related variants (human only). General goal of this study was to provide: i) information about the different organism-specific sequence datasets available from UniProt; and ii) the pros and cons of each one for bottom-up MS proteomics workflows. In addition, this study: - might help in annotating protein sequences into UniProt at the level of the "protein existence" field - can help in deciding what to import into UniProt from other DBs, when unique peptides present in other DBs carry substantial experimental evidence coming from the repositories - can serve for detecting potential proteotypic peptide candidates to be used in targeted proteomics workflows like Selected Reaction Monitoring

42

CSEO – The Cigarette Smoke Exposure Ontology

Erfan Younesi[1], Sam Ansari[2], Michaela Guendel[1], Shiva Ahmadi[1], Chris Coggins[3], Julia Hoeng[2], Martin Hofmann-Apitius[1], Manuel Peitsch[2]

[1] Fraunhofer Institute for Algorithms and Scientific Computing SCAI, Germany
[2] Philip Morris International R&D, Switzerland
[3] Carson Watts Consulting, United States of America

Presenter: Sam Ansari

In the past years, significant progress has been made in the development and use of settings for collection of experimental data on tobacco exposure and the diseases induced by it. Despite the growing number of such data, there has been no community wide effort to facilitate the centralization and integration of tobacco exposure data scattered throughout a range of disparate sources. Moreover, to fulfill the aim of exposure and disease impact studies, it is of utmost importance to more reliably and efficiently establish the causal link to disease. Ontologies are structural frameworks for organizing knowledge, enabling information retrieval, and supporting data integration, data analysis, and exchange of knowledge within the community. Cigarette Smoke Exposure Ontology (CSEO) was developed which is a specialized ontology with particular focus on the cigarette smoke exposure and related various experimental systems. Combining efforts of domain experts as well as novel computational methods, the ontology successfully describes exposure related terminology ranging from the scope and design definition of an experiment to its outcome with link to molecular events and ultimately to disease. After several iterations between the computational and domain expert groups, CSEO encompasses more than 20.000 concepts and classes. This ontology is represented in web ontology language (OWL) format and is made freely available to the community through several channels.

43

The Candida and Aspergillus Genome Databases: curated gene, protein, and genomic information resources for the fungal research community

Martha Arnaud[1], Jonathan Binkley[2], Gustavo Cerqueira[3], Diane Inglis[2], Marek Skrzypek[2], Prachi Shah[2], Farrell Wymore[2], Gail Binkley[2], Clinton Howarth[3], Stuart Miyasato[2], Matt Simison[2], Jennifer Russo Wortman[3], Gavin Sherlock[2]

[1] Aspergillus and Candida Genome Databases, United States of America
[2] Department of Genetics, Stanford University School of Medicine, United States of America
[3] Broad Institute, United Kingdom

Presenter: Martha Arnaud

The Aspergillus and Candida Genome Databases (AspGD, http://www.aspgd.org and CGD, http://www.candidagenome.org/) are freely available, web-based resources for researchers studying the molecular biology of these fungi. We have completed manual curation of the published literature about several Candida and Aspergillus species, and the interfaces of both web sites provide streamlined, ortholog-based navigation of the genomic and functional annotation for multiple species concurrently. AspGD also offers a full-featured genomics viewer to facilitate comparative genomics analysis. As part of our community-oriented mission, we also provide resources to foster interaction and dissemination of community information, tools, and data, including collecting, archiving, and providing large-scale datasets for download. At CGD, we recently undertook a project to improve Gene Ontology annotation for Candida biofilm formation, filamentous growth and phenotypic switching, which are traits of particular interest due to their association with virulence in these opportunistic pathogens. The project comprised three phases: first, the addition of terms to the Biological Process branch of the GO to improve the description of fungal-related processes; second, manual recuration of gene product annotations in CGD to use the improved GO vocabulary; and third, computational ortholog-based transfer of GO annotations from experimentally characterized gene products annotated with these new terms to uncharacterized orthologs in other Candida species. We welcome, encourage, and appreciate your questions, feedback or suggestions. AspGD and CGD curators can be reached at aspergillus-curator@lists.stanford.edu and candida-curator@lists.stanford.edu, respectively. AspGD is funded by grant R01 AI077599 from the National Institute of Allergy and Infectious Diseases; CGD is funded by R01 DE015873 from the National Institute of Dental and Craniofacial Research at the US National Institutes of Health.

44

Gene family-led curation at FlyBase.

Helen Attrill, Steven Marygold, Susan Tweedie, Nicholas H. Brown

FlyBase Cambridge, United Kingdom

Presenter: Helen Attrill

FlyBase is undertaking a review of gene-level data and creating a Gene Family resource for Drosophila melanogaster . Genes are grouped into well-defined families/groups (e.g. functional homologs, paralogs, multiprotein complex components) based on recent literature and the associated data in FlyBase are reviewed. This strategy has allowed us to make significant improvements to the consistency of gene ontology (GO) annotation and nomenclature, which was not being addressed by our usual paper-by-paper curation approach. To date, 64 gene families have been examined (representing 1214 genes - 7.6 % of total mapped genes), leading to the addition of 947 GO annotations. In addition, 196 were either replaced with a more appropriate GO term or removed because of erroneous annotations from high-throughput data or out-of-date gene models. Where possible, gene nomenclature has been unified, respecting current community usage. Additionally, this strategy has resulted in the removal or merging of 12 gene records and addition of 3 new genes to the database. The collection of gene groups will be used as a basis to form a Gene Family portal on the FlyBase website. This will provide an intuitive way to view members of gene groupings attributed to specific literature and easily access associated data (e.g. GO annotations, phenotypes). This interface will be particularly accessible to naive or non-drosophilist users of FlyBase and allow easy cross-reference with other gene family resources (such as those maintained by The HUGO Gene Nomenclature Committee (HGNC) and The Arabidopsis Information Resource (TAIR) databases).

45

A guide to best practices for Gene Ontology manual annotation

Rama Balakrishnan[4], Midori Harris[1], Rachael Huntley[2], Kimberley Vanauken[3], J. Michael Cherry[4]

[1] Pombase, United Kingdom
[2] UniProtKB, United Kingdom
[3] WormBase, United States of America
[4] Stanford University, United States of America

Presenter: Midori Harris

The Gene Ontology Consortium (GOC) is a community-based bioinformatics project that classifies gene product function through the use of comprehensive structured controlled vocabularies. A fundamental application of the Gene Ontology (GO) is in the creation of gene product annotations, evidence-based associations between GO definitions and published results. Currently, the GOC disseminates 126 million annotations covering more than 374 thousand species including all the kingdoms of life. This number includes two classes of GO annotations: those created manually by experienced biocurators reviewing the literature or expert analysis of biological data using specialized tools (1.1 million annotations covering 2226 species) and those prepared computationally via automated methods. Manual annotations are often used to propagate functional predictions between related proteins within and between genomes, therefore it is critical to have accurate manual annotations. Towards this goal we present here the conventions defined by the GOC for the creation of manual annotation. This guide represents the best practices for manual annotation that has been established by the GOC project over the past 12 years. We also hope this guide with encourage research communities to annotate gene products of their interest to enhance the corpus of GO annotations available to all.

46

Fusion Gene Curation in COSMIC

Sally Bamford, Charlotte Cole, Sari Ward, David Beare, Nidhi Bindal, Simon Forbes, John Gamble, Prasad Gunasekaran, Mingmin Jia, Chai Yin Kok, Kenric Leung, Frances Martin, Rebecca Shepherd, Jon W. Teague, P. Andrew Futreal, Michael Stratton, Peter J. Campbell

Wellcome Trust Sanger Institute, United Kingdom

Presenter: Sally Bamford

The Catalogue of Somatic Mutations in Cancer (COSMIC, http://www.sanger.ac.uk/cosmic) is an extensive public resource enabling the investigation of somatic mutations in human cancer. It includes mutation data curated from the scientific literature combined with sequencing data from several sources. The content now covers over 15327 manually curated publications. Since 2007 gene fusion have been added to the database to complement the data on point mutated cancer genes. COSMIC v62 (Nov 2012) describes almost 9000 gene fusions in 169 gene pairs. Gene fusions result from chromosomal rearrangements where typically an aberrant juxtaposition of 2 genes results in a fusion gene, and subsequent tumour-specific fusion protein with new or altered activity, or the regulatory elements of one gene may drive the aberrant expression of an oncogene. The Cancer Gene Census is dominated by fusion genes that have been identified in leukaemias, lymphomas and soft tissue tumours. The identification of TMPRSS2, a gene frequently found to be fused to ETS family transcription factors in adenocarcinoma of the prostate, stimulated the discovery of additional fusion genes in solid tumours. COSMIC aims to curate these newly identified fusions of solid tumors as well as the known fusion literature. One of the challenges has been the annotation and representation of these complex mutations and a curation tool has been developed to capture this mutation data at the varying levels of detail described in the literature. Fusions are curated at 2 levels: observed mRNA and genomic breakpoint. More than one mRNA transcript can be detected in an individual tumour sample and the genomic breakpoint is inferred based on all observed mRNAs. Fusions are not always straightforward recombinations and curation tools need to cope with a range of complexities at the breakpoint. Generation of appropriate syntax to describe fusions is also required.

47

Enzymes as drug targets: curated pharmacological information in the Guide to PHARMACOLOGY

Helen Benson[1], Elena Faccenda[1], Joanna Sharman[1], Adam Pawson[1], Doriano Fabbro[2], Daniel Treiber[3], Stephen Alexander[4], Michael Spedding[5], Anthony Harmar[1], NC-IUPHAR [6]

[1] The University/BHF Centre for Cardiovascular Science, United Kingdom
[2] Piqur Therapeutics, Switzerland
[3] KINOMEscan Division of Discoverx Corporation, United States of America
[4] School of Biomedical Sciences, United Kingdom
[5] Les laboratoires Servier, France
[6] The International Union of Basic and Clinical Pharmacology Committee on Receptor Nomenclature and Drug Classification, United Kingdom

Presenter: Helen Benson

Enzymes constitute a significant proportion of the druggable genome. In order to fully exploit their potential as drug targets it is vital that database tools exist to provide easily-navigable access to relevant genetic, biochemical and pharmacological information. The Guide to PHARMACOLOGY portal (http://www.guidetopharmacology.org/) is an open access resource providing expert-curated information on human drug targets and the substances that act on them. We have recently expanded the information available on enzyme drug targets, including the addition of >500 protein kinases and quantitative data on their interactions with 72 approved drugs and experimental inhibitors. The resource includes information on enzyme nomenclature, substrates, reactions, cofactors, inhibitors, links to relevant external resources and references, and expert overviews of their functions and (patho)physiology. For a subset of important targets we provide detailed expert-curated summaries from the primary literature on a wide range of properties. A pilot study to investigate the data types that needed to be added to the database for enzymes as drug targets involved curation of the enzymes of the lanosterol biosynthesis pathway, which includes the target of statin drugs used to treat hypercholesterolemia. This work formed a template for the curation of other enzymes such as the kinases, and is continuously adapted as new enzyme targets and data types are added. The Guide to PHARMACOLOGY now includes >1050 distinct enzymes, adding to existing information on >1200 receptors, ion channels and transporter proteins. This marks a significant milestone in our mission to provide expert-curated information for all the targets of current prescription medicines and other likely targets of future small molecule drugs via the Guide to PHARMACOLOGY portal. This is a unique resource which should appeal to scientists from a range of disciplines and aid further exploration of enzymes as drug targets.

48

Incorporating Ensembl Non-Coding RNA Genes into MGD

Sophia Zhu, Monica McAndrews, Dmitry Sitnikov, Janan Eppig, Judith Blake, Carol Bult

The Jackson Laboratory, United States of America

Presenter: Judith Blake

In the spring of 2013, the Mouse Genome Database (MGD) added 4500+ non-protein-coding Ensembl RNA genes to the database, more than doubling the number of ncRNA genes in MGD to over 8700. The cohort contains 983 micro RNA, 1462 small nucleolar RNA, 1421 small nuclear RNA, 157 ribosomal RNA, and 483 miscellaneous RNA genes. Genes are assigned marker and feature types upon importation. Meaningful nomenclature is given as each member of each RNA type is examined. Ribosomal RNAs are matched to the known ribosomal RNA groups, 5S, 5.8S, etc., given the appropriate root symbol and the next number in series. Small nucleolar RNAs may be categorized as Snora# or Snord#, depending on the presence of C/D box or H/ACA box sequences or Scarna#, small Cajal body-specific RNA. snRNAs are categorized by type of U-RNA U1, U2, etc, and given the next number in the series:Rnu1-#, Rnu2-#. Miscellaneous RNAs are blasted against known RNAs to aid in giving them nomenclature. We work with miRBase to give meaningful nomenclature to the microRNA genes. After chromosomal location is assigned, complex/cluster/region marker types can be created for microRNA clusters. The Gene Ontology (GO) group at MGI annotate miRNAs follwoing publication of functional and biological process analyses. MGI contains 735 GO annotations for 268 mouse miRNAs currently. MGD (http://www.informatics.jax.org) is part of Mouse Genome Informatics (MGI), the international resource for integrated genetic, genomic, and biological data about the laboratory mouse. Data in MGD are obtained through loads from major data providers and experimental consortia, and from biomedical literature. MGD maintains a comprehensive, unique catalog of mouse genome features generated by refining gene predictions from NCBI, Ensembl, and VEGA. MGD serves as the authoritative source for the nomenclature of mouse genes, mutations, alleles, and strains. MGD is funded by NIH NHGRI grant HG000330.

49

The M:N Project at MGD: Beyond 1:1 Orthology Assertions

Judith Blake, Richard Baldarelli, Mary Dolan, Mark Airey, Jonathan Beal, Sharon Giannatto, Dave Meirs, Jill Lewis, Carol Bult, Janan Eppig, James Kadin

The Jackson Laboratory, United States of America

Presenter: Judith Blake

The Mouse Genome Database (http://www.informatics.jax.org) curates, integrates, and provides comprehensive genetics, genomic, and phenotypic information for the laboratory mouse, a primary model organism for experimental investigation of human biology and disease. A core component of MGD data for over 20 years has been the curated assertion of 1:1 orthology between mouse, human, and rat protein-coding genes, work done in coordination with the human and rat genome annotation teams and gene nomenclature committees. Now, with comprehensively sequenced genomes available for comparative analysis, phylogenetic analysis clearly identifies cases where descent from common ancestor does not always define a 1:1 relationship, but rather that gene duplication following an ancestral speciation event more correctly results in M:N relationship between genes in different species. This has implications for the study of human biology in the mouse system and for the presentation of inferential functional and disease associated assertions based on comparative analysis. MGD has recently restructured its database to accommodate such homology classes with concurrent changes in presentation of data related to homology classes (primarily defined by NCBI HomoloGene resource) and in the representation of human diseases associated with mouse genes by curation of comparative or experimental data. We incorporate all the homology classes for mammalian species mouse, human, rat, chimp, cattle, dog and rhesus monkey, and will next extend our data to include chicken and zebrafish protein-coding gene classes. While 1:1 assertions predominate (~80%), we now more clearly represent cases such as the Serpina1 gene class (1 human, 5 mouse, 1 rat), and provide better crosss-referencing among related genes, the diseases that have been studied in respect to those genes, and the relationship between genomic features in related genomes. This work is supported by NIH NHGRI grant HG000330.

50

Catching inconsistencies in UniProtKB/Swiss-Prot with the semantic web

Jerven Bolleman, Sebastien Gehant, Nicole Redaschi

Swiss Institute of Bioinformatics, Switzerland

Presenter: Jerven Bolleman

The UniProtKB/Swiss-Prot database is manually curated by a team of experienced biocurators with the aim of providing the scientific community with high-quality information on proteins. Ensuring a high-quality curation standard depends in part on effective tools that help curators avoid trivial mistakes during data curation. We describe a system that uses SPARQL queries encoded in SPIN to identify UniProtKB database records that do not comply with manual curation rules. The system generates specific and accurate warnings for curators by correctly defining known exceptions to general rules.

51

PDBWiki: success or failure? Three rules for successful community annotation projects

Dan Bolser[1], Jose Duarte[2], Jong Bhak[3], Henning Stehr[4]

[1] EMBL-EBI, United Kingdom
[2] Paul Scherrer Institut, United Kingdom
[3] Theragen BiO Institute, Republic of Korea
[4] Max Planck Institute for Molecular Genetics, Germany

Presenter: Dan Bolser

Community annotation of biological data is of increasing importance for several reasons. As the cost of biological data decreases, the volume of data available to ever more specialised communities is growing superlinearly. As a consequence, the growing task of data annotation is falling to ever fewer experts. The success of the world's largest community annotation experiment, Wikipedia, suggests a possible solution to this problem. But, what factors contribute to a successful community annotation project? We created PDBWiki, a biowiki for community annotation of the macromolecular structures in the Protein Data Bank (PDB). Initially, this project met our need for a way to share information on problematic protein structures. As the wiki developed, however, it allowed us to explore the issues surrounding community annotation of biological data. Based on our experience and drawing from the literature, we have identified three interrelated rules, or guidelines, for successful community annotation projects. 1) Functionality: the project must provide a useful resource from the beginning. 2) Benefit: there should be some immediate reasons, implicit or explicit, for users to contribute. 3) Recognition: in the long-term, user contribution should be rewarded, either by publication or using principles of social engineering.

52

GUDMAP: GenitoUrinary Development Molecular Anatomy Project

Jane Brennan[1], Jane Armstrong[1], Sue Lloyd MacGilp[1], Simon Harding[2], Bernie Haggerty[2], Yogmatee Roochun[2], Jamie Davies[1], Richard Baldock[1]

[1] Edinburgh University, United Kingdom
[2] IGMM, University of Edinburgh, United Kingdom

Presenter: Jane Brennan

The GenitoUrinary Development Molecular Anatomy Project (GUDMAP) is a consortium of NIH-funded laboratories working to provide the scientific and medical community with gene expression data, transgenic mice and tools to facilitate research. The GUDMAP Editorial Office and Database Development Team – both based in Edinburgh – function to ensure submission, curation, storage and presentation of the data submitted by the GUDMAP consortium. The data is housed in the GUDMAP-database (www.gudmap.org), which now includes over 10 000 in-situ hybridisation entries and over 400 microarray samples of microdissected, laser-captured and FACS-sorted components of the developing mouse genitourinary (GU) system. The expression data is annotated using a high-resolution ontology specific to the developing murine GU system and is freely accessible via easy-to-use interfaces.

53

Manual curation of the human proteome. Status and collaborations.

Lionel Breuza[1], SIB Swiss Institute of Bioinformatics Swiss-Prot Group[1], The European Bioinformatics Institute The EMBL Outstation[2], Protein Information Resource[3]

[1] SIB Swiss Institute of Bioinformatics, Switzerland
[2] The European Bioinformatics Institute, United Kingdom
[3] Protein Information Resource, United States of America

Presenter: Lionel Breuza

The UniProt Knowledgebase (UniProtKB) provides the scientific community with a stable, comprehensive, and richly annotated resource of protein sequences and functional information. It is composed of a manually curated section, UniProtKB/Swiss-Prot, and an automatically annotated complement, UniProtKB/TrEMBL. The manual curation of the human proteome is a priority of the consortium and includes the integration of information extracted from the literature and the thorough analysis of protein sequences. All human protein-coding genes have been manually reviewed and are described within 20,226 UniProtKB/Swiss-Prot records, including over 15,000 reviewed alternatively spliced sequences, while more than 50,000 unreviewed isoform sequences are available in UniProtKB/TrEMBL (statistics for release 2013_01). We continually revisit human UniProtKB/Swiss-Prot entries as knowledge evolves, updating functional annotation based on newly published primary literature, and our manual curation workflow includes the routine assignment of Gene Ontology (GO) terms to human proteins. We also update sequences and associated annotation, curating new variants, their functional impact and association with diseases, new alternatively spliced isoforms, and reviewing and correcting erroneous sequences. These sequence curation activities are performed collaboratively with other resources including HAVANA, Ensembl, RefSeq and the Consensus CoDing sequence project (CCDS), and we regularly revise gene model annotations in response to feedback from these resources in order to improve their quality and consistency.

54

Proteome set reorganization in UniProtKB

Maria-Jesus Martin, Ramona Britto

EMBL-EBI, United Kingdom

Presenter: Ramona Britto

The Universal Protein Resource (UniProt) endeavours to support biological research by providing a stable, comprehensive, consistent and accurately annotated protein knowledgebase that is freely accessible by the scientific community. In addition to standardizing and integrating data from numerous resources, UniProt provides rich and comprehensive functional annotation of its protein sequences. As the cost of sequencing continues to fall, there are an increasing number of cases where the same organism or subspecies has been completely sequenced by more than one project. The multiplicity of resulting proteomes has led to new challenges for UniProtKB necessitating a reorganization of proteomes in the database. Historically, nuclear and organellar genomes were sequenced independently; for the purpose of complete proteome presentation in UniProt, these are brought together via the NCBI taxonomy identifier (taxid). As this taxid is no longer sufficient to distinguish protein sets from different genome assemblies, a new unique proteome identifier will be introduced to group together proteome components (e.g. chromosomes, whole genome sequencing contigs and organelles). We hope, in the long term, that the new identifier will unambiguously identify the set of components that make up each proteome. Also in the pipeline are measures to reduce redundancy within UniProt proteomes, in particular, the implementation of a gene-centric view of proteome sets. This approach will group together canonical and variant protein sequences and provide a representative protein for each gene in the genome. This poster discusses the introduction of a new proteome identifier to group together nuclear and organellar proteomes pertaining to a single species and to help distinguish between multiple sequencing projects for the same organism.

55

GenomeRNAi Database: Curation and Data Annotation of RNA Interference Screens

Svetlana Buhlmann, Esther Schmidt, Oliver Pelz, Grainne Kerr, Arunraj Dhamodaran, Klaus Yserentant, Michael Boutros

German Cancer Research Center (DKFZ), Division Signaling and Functional Genomics, United Kingdom

Presenter: Svetlana Buhlmann

RNA interference (RNAi) is a powerful method to investigate gene loss-of-function. Systematic large-scale and genome-wide RNAi screens generate phenotypes in a variety of organisms as a source of functional gene annotation. The publicly accessible database GenomeRNAi (www.genomernai.org) aims to collect and make available published (manually curated) and author-submitted RNAi phenotype data and to facilitate comparison across multiple screening experiments. It also provides detailed information on approx. 460 000 RNAi reagents, e.g. calculations on specificity and efficiency. The intuitive web-interface allows to search by gene, reagent or phenotype or to browse through RNAi screens or frequent hitters (genes that frequently show a phenotype). Furthermore, data can be downloaded in TXT format or visualised in a genome browser via a DAS server. Currently, GenomeRNAi comprises information from 137 cell-based experiments in Homo sapiens , as well as 173 screens in Drosophila melanogaster , 53 of which were performed in vivo . With increasing numbers of published RNAi-induced phenotypes, the systematic integration and annotation of functional information remains a major challenge. Considering the wide variety of biological assays applied in cell-based and in vivo RNAi screening experiments, annotation guidelines are essential for conclusive comparisons across various datasets. Since no phenotypic guide for the description of RNAi experiments is available, we defined controlled vocabularies and postulated vocabulary guidelines for annotation of RNAi screens and phenotypes. Moreover, we developed a pipeline for curation, annotation and database insertion comprising data quality controls. We will report on curatorial and annotation criteria and guidelines, as well as curation progress and current database content.

56

Using GMOD to annotate and distribute data in the cloud

Scott Cain[1], Amelia Ireland[2]

[1] Ontario Institute for Cancer Research, United States of America
[2] University of California at Berkeley, United States of America

Presenter: Scott Cain

As sequencing technology becomes faster and less expensive, there is an increasing need for computing resources and software to process the data it produces. Making use of "cloud computing"--hosting data and/or applications on existing networked computer systems--to give users access to preconfigured, extensible servers is an alternative to building and maintaining large computing infrastructure in-house. The GMOD project has created a community annotation server with a Chado database, GBrowse2, JBrowse, Tripal, and WebApollo installed, hosted by Amazon Web Services. Users can clone the GMOD Amazon Machine Image (AMI) and create their own server for storing data and making it accessible to the public. Potential applications of the GMOD AMI range from short term usage, such as during annotation jamborees, to the long term provision of access to genome data and applications to the community.

57

AmiGO 2: An Adventure in Flatland

Seth Carbon

Lawrence Berkeley National Lab, United States of America

Presenter: Seth Carbon

AmiGO 2 is an open-source web application in development for the Gene Ontology that allows users to query, browse, and visualize ontologies and related gene product annotation data. By intelligently applying a document-oriented approach to the problems that are encountered when dealing with ontology and annotation data, AmiGO 2 has greatly increased its performance for common tasks, as well as expanding its feature and development space. AmiGO was initially built around a relational database (RDB), but over time performance and functionality issues have increasingly hindered progress (e.g. overly complicated graph queries and ranking search data quickly). To overcome these issues, AmiGO 2 is built on a Solr document store; and while document stores have an inherently less rich structure than RDBs, being essentially restricted to fields and values in individual documents, intelligent field creation, planning, and leveraging the nature of document stores can cover most of the use cases of previously covered by RDBs and related software. In the case of AmiGO 2, the result has been greatly increased speed, simplicity in loading, and faster development. Moreover, previously expensive and time consuming problems have become quite easy--complicated searches involving both text matching, filtering, and ontology closures can be performed in a single instantaneous step and data exploration can be easily accomplished with facets. Abilities like these allow for the creation of richer and faster tools to explore data in more layered ways than before. With this, new opportunities abound. We can now provide: direct access to the data store, widgets that can be easily incorporated into any web page, JavaScript APIs for users to create their own applications, and scripting functionality for working with data with outside of a web browser. Our goal is to have these services and APIs used by any group interested in working with the Gene Ontology.

58

The European Nucleotide Archive (ENA): Genome assemblies and complete genomes submission process

Ana Cerdeno-Tarraga, European Nucleotide Archive

EMBL-EBI, United Kingdom

Presenter: Ana Cerdeno-Tarraga

The European Nucleotide Archive (ENA) objective is to support and promote the use of nucleotide sequencing as an experimental research platform by providing data submission tools via a variety of routes, archive, search and download services. The increasing number of sequence submissions from small-scale as well as large-scale genomic sequencing projects has required ENA the development of concepts and standards for curation and data validation as well as automated tools that assist both the submitter and the biocurator. ENA’s new submission system (Webin)1 allows submitters to select from a list containing mandatory and optional fields based on the type of data to be submitted, and uses a biological rule-based validator. For large-scale data submissions this system includes three validated options: one for complete genome data in an EMBL-formatted file, one for annotated Whole Genome Shotgun (WGS) and one pre-formatted check-list for unannotated WGS. Large sequencing centres however, require a more automated way of submission and so they use validated direct pipelines for their large-scale data submission to ENA. In order to comply with INSDC standards, submitters of complete genomes or genome assemblies are requested to register their large-scale project with ENA and to provide metadata related to the project’s description, assembly and sequence. This information can be easily uploaded into their own private ftp directory for quick processing. Here we present examples of our large-scale data submission system options, assembly metadata required and checks that have been implemented into the ENA’s validator. 1-Major submissions tool developments at the European Nucleotide Archive. Amid et al. NAR, 2012;40:D43-7.

59

Next generation pfsearch: accelerated search of PROSITE generalized profiles

Lorenzo Cerutti, Thierry Schuepbach, Marco Pagni, Alan Bridge, Lydie Bougueleret, Ioannis Xenarios

SIB Swiss Institute of Bioinformatics, Switzerland

Presenter: Lorenzo Cerutti

The PROSITE resource provides a rich and well annotated source of signatures in the form of generalized profiles that allow protein domain detection and functional annotation. While PROSITE has been used to identify and annotate functional domains during the manual curation of UniProtKB/Swiss-Prot for over a decade now, its application in genome and metagenome annotation pipelines has been limited by the length of time required to search large protein sequence databases for putative matches. To address this issue, we have developed an optimized implementation of the PROSITE search tool pfsearch as well as a new heuristic that selects potential candidate sequences for a full sequence-profile alignment. On modern computers our new implementation of pfsearch is 2 orders of magnitude faster than the original, facilitating the application of PROSITE to very large datasets.

60

Annotating protein sequences – what InterPro can do for you!

Alex Mitchell, Amaia Sangrador-Vegas, Siew-Yit Yong, Sarah Hunter, Hsin-Yu Chang

EBI-EMBL, United Kingdom

Presenter: Alex Mitchell

InterPro (www.ebi.ac.uk/interpro) is a protein sequence analysis resource that brings together 11 different protein signature databases. Our main goal is to provide an accurate and easy way to classify proteins into families and identify domains and sites. So far, we have manually integrated entries that cover around 80% of the UniProtKB sequences. To help our users to comprehend their analysis results, descriptive abstracts, references and suitable Gene Ontology terms are added to each entry by our curators. Our resource is often used for automated analysis in large scale sequencing projects, such as the analysis of whole proteomes, genomes and metagenomes. It is also used for small scale analyses, providing detailed characterisation of individual protein sequences. In this poster, we show how InterPro can help scientists and curators to annotate protein sequences, explore protein families and gather information across different species.

61

Sequence Submissions to UniProt Knowledgebase (UniProtKB)

Gayatri Chavali[1], Consortium UniProt[2]

[1] EMBL-EBI, United Kingdom
[2] EMBL-EBI, Swiss Institute of Bioinformatics, Protein Information Resource

Presenter: Gayatri Chavali

The UniProt Knowledgebase (UniProtKB) is a centralised resource of protein knowledge that provides comprehensive and authoritative protein sequence and functional information (http://www.uniprot.org). Manual and automatic annotation procedures are used to add data directly to the database. In addition, extensive cross-referencing with more than 120 external databases provides access to additional scientifically relevant information in more specialised contexts. The data is freely available in a range of formats to facilitate integration with other databases. UniProtKB accepts depositions and issues accession numbers for directly sequenced proteins. Direct data submissions are accepted from the research community via SPIN (http://www.ebi.ac.uk/swissprot/Submissions/spin/index.jsp), -the web-based interactive submission tool. SPIN submissions are annotated using data provided by submitters combined with results from sequence analysis tools and information propagated from homologous proteins in the database. While most submissions contain a small number of sequences, bulk submissions are also accepted and work on SPIN is on-going to improve the bulk submission process. Submissions include both published and unpublished data. Published data is processed for immediate release while unpublished data can be held confidential by depositor request until publication. The current procedures detailing the annotation of direct data submissions are described.

62

Manual curation of Drosophila melanogaster in UniProtKB

Elena Cibrian-Uhalte[1], UniProt Consortium[2]

[1] EMBL-EBI, United Kingdom
[2] UniProt, United Kingdom

Presenter: Elena Cibrian-Uhalte

Drosophila melanogaster has been used as a model organism for genetic and developmental studies for over a century. The publication of its complete genome sequence in 2000 broadened its use for the study of human diseases so that today there are fly models for a wide range of conditions, including cancer and neurological disorders. Moreover, recent genetic analyses have shown that around 75% of human disease genes have homologs in Drosophila melanogaster. Due to its relevance in biomedical research, Drosophila melanogaster is a priority model organism for UniProtKB manual curation. The Drosophila annotation program focuses on the manual curation of characterised Drosophila melanogaster proteins with experimental data from the scientific literature. Since 2006, all Drosophila melanogaster entries in UniProtKB contain the keyword 'Complete proteome', allowing the retrieval of a complete non-redundant set of proteins based on the translation of the Drosophila melanogaster completely sequenced genome, which currently comprises 17,533 entries and which can be downloaded from the UniProt website. All Drosophila melanogaster entries are cross-referenced with FlyBase, the main database for Drosophila genetics and molecular biology, with which UniProtKB keeps close contact in order to ensure data consistency between both databases. In addition to melanogaster, 11 other Drosophila species are currently present in UniProtKB as complete proteome sets. Functional annotation of these Drosophila species is achieved by propagation of experimental information from D. melanogaster orthologous proteins.

63

27 and rising : Improving Pfam

Penelope Coggill[1], Ruth EBerhardt[1], Robert Finn[2], Jaina Mistry[1], John Tate[1], Alex Bateman[1], Marco Punta[1]

[1] EMBL-EBI, United Kingdom
[2] HHMI Janelia Farm Research Campus, United States of America

Presenter: Penelope Coggill

Pfam aims to provide a complete and accurate classification of protein sequences. Here we provide an overview of progress towards this goal. The latest release of Pfam (version 27.0) is based on sequences from UniProtKB and contains 46% more proteins than release 26.0. Between releases 26 and 27 we have added 1158 more families and, despite the large increase in the number of proteins, sequence- and residue-coverage - our progress metrics - have not significantly altered. Most of the recently built families are relatively small suggesting we have the ubiquitous families (100,000s of matches) and there are unlikely to be large areas of protein-space containing sets of, previously unidentified, homologous proteins. While a number of uncovered proteins may be genus- or species-specific, not all remaining proteins are. Some sequences/residues lie just outside the boundaries of existing families, and some contain sequence-compositions not handled by current homology-based methods, that can lead to significant non-homologous matches. All these remain a challenge for curators, requiring intense manual curation to avoid mis-annotation and false positives. Algorithmic developments are in-hand to enable HMMER to improve sensitivity for regions of lower similarity such as at the edges of domains and for motifs and short repeat regions. Visualising all sequences in very large families has become a major challenge. For Pfam release 27, we will offer multiple sequence alignments based on four progressively non-redundant sets of representative complete proteomes. This release will also offer the option to align or download sequences from any segment of the Sunburst representation of the species-trees. We have also refined our DNA-search capability, and will provide a full 6-frame translation run against the database.

64

Sequence Curation of Reference Assemblies by the Genome Reference Consortium

Joanna Collins

Wellcome Trust Sanger Institute, United Kingdom

Presenter: Joanna Collins

The Genome Reference Consortium (GRC) was established with the aim of updating and improving the reference assemblies for Human, Mouse and Zebrafish genomes to create new assemblies that better represent allelic diversity and provide more robust substrates for genome analysis. In addition to representation of variant loci curators are responsible for providing ‘genomic care’ for chromosomes in order to improve the current reference assemblies. Examples of this work include resolution of alignment discrepancies with transcripts, identification and repair of misassembled sequence, gap closure and retiling of problematic regions. Assembly improvement is achieved using standardised operating procedures. Issues requiring assessment are reported either by the partners, or the scientific community, resulting in curators assuming responsibility for the management of these issues. A variety of genome analysis tools are employed by curators in order to reach a satisfactory resolution to the issues raised, these can be classified as Tracking systems, Sequence evaluators and Genome Browsers. Examples of those used include JIRA, GenomeWorkbench, PGPviewer and Ensembl. Regions under review and progress are reported at genomereference.org. Improved/corrected reference assembly sequence is accessible to the research community through the release of genome patches in minor assembly updates between major assembly releases. Patch scaffolds have an alignment to the primary assembly but at the time of release are not integrated into the reference chromosomes and therefore do not disrupt chromosome coordinates. There are two types of genome patch: Fix patches, providing a temporary representation of the corrected sequence to be merged into the next major release, and Novel patches, adding a permanent variant of an existing region. Examples of both Fix and Novel patches, alongside a description of the curation tools that are integral to their development will be presented.

65

Virtual Fly Brain

Marta Costa[1], David Osumi-Sutherland[1], Simon Reeve[1], Nestor Milyaev[2], Cahir O'Kane[1], J. Douglas Armstrong[2]

[1] Department of Genetics, University of Cambridge, United Kingdom
[2] University of Edinburgh, School of Informatics, Institute for Adaptive and Neural Computation, United Kingdom

Presenter: Marta Costa

Navigating the Drosophila neurobiology literature and related databases is challenging. For example, it can be a daunting task to find details of the connectivity between two brain regions, the properties of the neurons involved, and the genes and GAL4 drivers that they express. This problem is rapidly growing worse as yet larger datasets are produced. One way to tie neuroanatomical data together is in an atlas. Google Earth, with its ability to rotate, zoom, overlay data and link any feature to additional information is an obvious template. Inspired by this approach, we have developed the Virtual Fly Brain (VFB), a web-based tool that allows users to browse a 3D confocal stack of a Drosophila brain at any angle and various scales. For any brain region down to the level of individual glomeruli and layers, users can run point-and-click queries for neuron classes based on innervation patterns, for alleles based on phenotype and for markers and GAL4 drivers based on expression. Subdivision of the brain on VFB is defined using names, boundaries and textual definitions agreed by the BrainName project [Ito et al., in preparation]. Annotations are stored in the FlyBase Drosophila anatomy ontology, which also stores detailed information from the literature about neuron classes, including their lineage, innervation patterns and neurotransmitters. This information can be searched on VFB via simple template-based queries for neuron classes. Phenotype and expression data is pulled directly from FlyBase, who use this ontology extensively in their curation. We are currently extending the data sets annotated with our ontology to other neuroanatomical resources. We are also incorporating alignment tools that allow users to register their stacks to our painted atlas stack and to annotate them using our ontology. With this approach, we aim to make VFB a hub for querying across multiple neuroanatomical resources and integrating them with genomic and literature resources.

66

NBLAST: a new tool to efficiently annotate large neuron image datasets

Marta Costa[1], David Osumi-Sutherland[1], Gregory Jefferis[2]

[1] Department of Genetics, University of Cambridge, United Kingdom
[2] Division of Neurobiology, MRC Laboratory of Molecular Biology,, United Kingdom

Presenter: Marta Costa

The 100,000 neurons in the adult Drosophila brain can be grouped into an estimated 5-10,000 classes. The Drosophila anatomy ontology has information on nearly 600 of these, all of which can be browsed and queried on Virtual Fly Brain (VFB; www.virtualflybrain.org). Many recent papers describe and analyse the function of new or known neuron classes in their efforts to map the neuronal circuits involved in coordinating behaviour. Ever-improving molecular tools that allow labelling of neurons have resulted in an increasing number of large datasets of single neuron images (e.g. flycircuit.tw) or transgene expression images being released to the community. The manual effort involved in curating these large datasets has resulted in only limited and ad hoc mappings of images to know neuronal classes being published, and there are no comprehensive analyses predicting new classes from such image sets. The challenge of making this data searchable includes not only the annotation of huge numbers of images, but also the ability to identify the same neuron class in different images. We have developed a tool that facilitates the annotation of large image datasets. NBLAST measures the similarity between pairs of neurons by location and morphology. A distance matrix of NBLAST scores for all pairs of neurons in a database can then be passed to a clustering algorithm, producing morphologically similar clusters. Given the close relationship between neuronal morphology and function, these clusters will likely correspond to neuron classes, known or yet unidentified. Using NBLAST, we are in the process of analysing and annotating ~16000 neuron images from FlyCircuit and incorporating the images and annotations into Virtual Fly Brain. We will present the preliminary results of this analysis for antennal lobe projection neurons, demonstrating that the neuron clusters generated by NBLAST match known neuron classes and so can be used to massively expand the mapped image content on VFB.

67

New Regulation Data in the Saccharomyces Genome Database.

Maria Costanzo[1], Stacia Engel[1], Gail Binkley[1], Esther Chan[1], Benjamin Hitz[2], Kalpana Karra[1], Paul Lloyd[1], Greg Roe[1], Shuai Weng[1], Edith Wong[1], J. Michael Cherry[1]

[1] Stanford University, United States of America
[2] SGD, United States of America

Presenter: Maria Costanzo

The Saccharomyces Genome Database (SGD; http://www.yeastgenome.org) is the community resource for genomic, gene, and protein information about the budding yeast Saccharomyces cerevisiae . SGD biocurators extract from the published literature a variety of functional information about each yeast gene and gene product. In addition to the Gene Descriptions, Gene Ontology annotations, mutant phenotype annotations, and other gene-specific data currently in SGD, we are preparing to add a new category, regulatory information. In the first phase of this project, we are focusing on 144 DNA-binding transcription factors (TFs). The Locus Summary page for each of these regulators will have a new tabbed section, Regulation, that will provide several different types of information about each TF. A short Regulation Summary Paragraph will give an overview of the biological context in which the TF acts. The consensus binding site motif sequences and logos will be displayed, and numbers of predicted and experimental binding sites will be summarized. The genomic distribution of binding sites will be displayed in a Circos genome visualization map. A table of targets of each regulator along with references and evidence, compiled from information curated at SGD and at the YEASTRACT database, will be displayed along with a summary of the biological processes in which the target genes are involved. The Regulation tab will also display the structural classification of each TF, and information regarding its regulation by other TFs. All of these data will be available for querying, analysis, and download via YeastMine, the Intermine-based data warehouse system in use at SGD. In the future, we plan to expand this information to include additional types of regulation. We thank the YEASTRACT group for permission to display their curated regulatory relationships in SGD. This work is supported by a grant from the US National Human Genome Research Institute (P41 HG001315).

68

The annotation resources and gene ontologies provided by the HAVANA group.

Claire Davidson, Jonathan Mudge, Jennifer Harrow

Wellcome Trust Sanger Institute, United Kingdom

Presenter: Claire Davidson

The HAVANA group produce comprehensive manual annotation of gene features for vertebrate reference genomes. Annotated models are classified as one of three biotypes: protein-coding, pseudogene or non-coding. The manual annotation process allows us to further sub-categorise our models, significantly increasing the information content of our build. Thus protein-coding transcripts are classed as known, novel or putative to reflect our level of confidence in the annotated coding sequence (CDS), pseudogenes are described as processed, unprocessed, unitary or polymorphic to indicate their mode of formation, while non-coding transcripts are categorised as antisense, lincRNA or sense intronic in order to capture their spatial relationships with protein coding loci. A third descriptive layer may be provided by attaching ontological 'attributes' to a model. Such attributes may (for example) highlight models containing non-canonical splicing motifs, and whether such sites are conserved in other species, or represent known polymorphisms. CDS-linked attributes may demarcate the presence of an ORF upstream of the CDS or the use of a non-canonical initiation codon. This approach allows us to generate genesets of incomparable richness, and is particularly beneficial in capturing the functional complexity of alternative splicing. Furthermore, manual annotation is essential for describing loci that prove problematic for automatic pipelines, such as complex gene families and pseudogenisation events. All HAVANA annotation is freely available and can be viewed in the VEGA genome browser. Human VEGA is now updated weekly, allowing us to rapidly share improvements made. The human GENCODE geneset, comprising a merge of HAVANA models and computationally derived Ensembl models can also be visualised via the Ensembl and UCSC genome browsers, and downloaded from gencodegenes.org.

69

Annotation of the Brugia malayi genome in WormBase

Michael Paulini[1], Kevin Howe[1], John Spieth[2], Philip Ozersky[2], Gary Williams[1], Paul Davis[1]

[1] EMBL - EBI, United Kingdom
[2] The Genome Institute, United States of America

Presenter: Paul Davis

Brugia malayi is a parasitic nematode that is one of the causative agents of lymphatic filariasis in humans (commonly known as elephantiasis). More than 1.3 billion people in 72 countries worldwide are threatened by lymphatic filariasis with over 120 million people currently infected(1). WormBase has collaborated with the University of Pittsburgh on the annotation of a much improved version of the reference genome sequence (first published in 2007). After predicting initial gene models using an automated RNASeq-guided pipeline developed in-house, we embarked on a pilot study to manually assess and curate 15% of the predicted Brugia malayi gene loci. By soliciting input from the Brugia malayi research community, curation efforts were directed towards genes of experimental interest, for example potential drug targets, protein kinases, and transcription factors. In addition, methods and procedures developed by WormBase for Caenorhabditis elegans curation were used to maximise the utility of the available Brugia malayi data and enable transfer of knowledge from the Caenorhabditis elegans curation effort. Numerous improvements were made to the existing curation tools, enabling the rapid curation of large numbers of loci in a relatively short amount of time. The curation model developed in this project will act as a template for future WormBase involvement in the annotation of parasitic nematode genomes.

70

Dealing with Proteomics large-scale mass spectrometry experiments in UniProtKB/Swiss-Prot

Edouard de Castro[1], UniProt consortium[2]

[1] SIB, Switzerland
[2] UniProt consortium, Switzerland

Presenter: Edouard de Castro

Results from high-throughput mass spectrometry-based proteomics experiments constitute a useful source of information for protein knowledgebases such as UniProtKB. However, efforts to capture this information are complicated by the continuous evolution of experimental technologies and the heterogeneous quality of peptide and protein identifications as well as associated experimental metadata. We have developed a flexible pipeline for the selection and integration of proteomics data that is of suitable quality for the annotation of UniProtKB/Swiss-Prot records. Experimental peptides, including those with post-translational modifications (PTMs), are extracted from proteomics reports according to manually curated criteria and are then incorporated into a unified database. These peptide sequences are then checked for their ”unicity”: peptides are considered unique if they map uniquely to one or more protein products (isoforms) from a single gene or to a group of identical (and therefore indistinguishable) proteins, which are the products of different genes. Protein entries identified in this way are categorized with ‘Evidence at protein level’ and annotated with the corresponding information on PTMs and derived annotations (such as keywords and literature citations), as are homologous protein sequences in closely related species. The annotation pipeline is compatible with the ProteomeXchange workflow, allowing information both from classical publications and well annotated datasets to be incorporated with appropriate tracking and documentation. The system provides full traceability of the source information for curators. The update of existing annotations according to changes in the selection and filtering criteria, and the addition of data from new publications can be easily performed. The pipeline has been applied to UniProtKB/Swiss-Prot protein sequence records of human origin and will be extended to other species in future releases.

71

PortEco: portal for E. coli research

Janos Demeter

Stanford University, United Kingdom

Presenter: Janos Demeter

E. coli is one of the best studied organisms on earth, but relevant information is scattered across the web. PortEco (http://www.porteco.org/) is designed to bring it all together for users of E. coli genomic data. The features of PortEco currently include: PortEco search, which simultaneously searches 13 different E. coli data sources, aggregating and organizing the results on a single results page; PortEco:Expression data repository, which stores and extensively processes expression datasets from many different sources, including tools for analyzing and interpreting data (currently mRNA microarray data); community features including news and blog aggregator, forum, and calendar; EcoliHouse, a database with E. coli information to support bioinformatics queries and EcoliWiki, community-contributed content about E. coli. The PortEco:Expression site holds post-publication microarray gene expression data from E. coli experiments. Currently it includes 125 datasets and our goal is to provide free access to all published E. coli expression data. PortEco expression uses software from SMD to visualize expression data. Additional tools were developed to visualize data for organism specific site. Some of these tools are shown, including Advanced Search, Samples and Conditions tool, Cluster My Genes as well as Gene Profiles.

72

Conserved Domain Database (CDD): automated procedures for the detection of conserved domain subfamilies; extending coverage of the protein domain universe.

Myra Derbyshire[1], Marc Gwadz[1], Christopher Lanczycki[1], Farideh Chitsaz[1], Noreen Gonzales[1], Fu Lu[1], Gabriele Marchler[1], James Song[1], Narmada Thanki[1], Roxanne Yamashita[1], Chanjuan Zheng[1], Stephen Bryant[1], Aron Marchler-Bauer[1], Andrew Neuwald[2]

[1] National Center for Biotechnology Information, United States of America
[2] National Center for Biotechnology Information and University of Maryland, School of Medicine, United States of America

Presenter: Myra Derbyshire

NCBI's Conserved Domain Database (CDD) is a protein classification and annotation resource. It contains a collection of multiple sequence alignments (MSAs) of protein domains that are being organized into hierarchical classification systems, based on conserved and divergent sequence and structural features. Manual curation of hierarchical classifications is time consuming and may be difficult to reproduce. While manual intervention is still necessary to address very large and diverse domain families, we are proposing to employ automated procedures for the detection of conserved domain subfamilies in the majority of all cases. Our goal is to approach comprehensive coverage of the protein universe. Recently, Neuwald AF, Lanczycki CJ, Marchler-Bauer A 2012, developed an automated procedure to rapidly generate hierarchical classifications of protein domain subfamilies. This method starts from a (typically very large) MSA, uses a combination of heuristic and Markov chain Monte Carlo sampling procedures, and is based on functionally-divergent residue signatures. It has been further optimized alongside the manual curation of new and updated CDD- hierarchies as CDD curators compare the output of the sampler with their results. We are in the process of sampling subfamilies for a large number of domain families, and will discuss our progress in integrating this procedure into the curation pipeline, as well as illustrate with examples from the curation of the globin and SDR superfamilies. This work was supported by the Intramural Research Program of the National Library of Medicine at the National Institutes of Health/DHHS and (to AFN) through an NCBI contract, the University of Maryland, and NIH grant GM078541

73

Hierarchical Orthologous Groups: What are they? How to infer them? Why use them?

Christophe Dessimoz

EMBL-EBI, United Kingdom

Presenter: Christophe Dessimoz

Hierarchical orthologous groups are defined as sets of genes that have descended from a single common ancestor within a taxonomic range of interest. Identifying such groups is useful in a wide range of contexts, but in particular they are of high relevance for accurate inference of gene function. Hierarchical orthologous groups can be derived from reconciled gene/species trees but, this being a computationally costly procedure, many phylogenomic databases work on the basis of pairwise gene comparisons instead (“graph-based” approach). In this talk or poster, I will expose correspondences between reconciled trees and the orthology graph. Based on these, I will briefly outline a novel inference method for hierarchical orthologous groups called GETHOGs ("Graph-based Efficient Technique for Hierarchical Orthologous Groups"). Using specific examples, I will illustrate how hierarchical groups can shed light on the evolution of complex gene histories and how they can provide a way to pinpoint the emergence of new gene function. References: Altenhoff AM, Gil M, Gonnet GH, Dessimoz C (2013) Inferring hierarchical orthologous groups from orthologous gene pairs. PLoS ONE 8: e53786. doi:10.1371/journal.pone.0053786. Boeckmann B, Robinson-Rechavi M, Xenarios I, Dessimoz C (2011) Conceptual framework and pilot study to benchmark phylogenomic databases based on reference gene trees. Brief Bioinform 12: 423–435. doi:10.1093/bib/bbr034.

74

Curating high quality intrinsic protein disorder annotations for the UniProt database

Tomas Di Domenico, Silvio Tosatto

University of Padova, Italy

Presenter: Tomas Di Domenico

During the last few years, intrinsic protein disorder has become an increasingly important topic in protein science. Due to the difficulty of experimentally characterizing the phenomenon, in silico predictions have been the main source of information used by the community. This situation is currently changing, and experimental determination of disorder is becoming increasingly feasible. So far, there have been several attempts at building databases to store intrinsic disorder annotations. Each of these have tackled the problem with different approaches. Here we present the latest version of the MobiDB database. The goal of this new version is twofold. The first objective is to provide the best possible disorder annotations for all of UniProt's close to thirty million proteins by leveraging existing data sources. The second, to provide a curation workflow to encourage the community to deposit experimental disorder information. The result can be thought of as a data pyramid. The base of the pyramid is composed of all proteins, which would feature the comparatively less reliable predicted annotations. The middle part of the pyramid would be populated by the subset of proteins which also feature some indirect evidence of disorder, like PDB structures. Finally, the top of the pyramid would consist on an ever smaller but very high quality set of proteins which, apart from featuring the aforementioned predictions and indirect evidence, have manually curated and reviewed intrinsic disorder annotations. The MobiDB database can be accessed via a user interface or web services, and it's freely available at http://mobidb.bio.unipd.it.

75

The banana genome hub

Gaëtan Droc[1], Delphine Lariviere[1], Valentin Guignon[2], Nabila Yahiaoui[1], Dominique This[3], Alexis Dereeper[4], Chantal Hamelin[1], Xavier Argout[1], Jean-François Dufayard[1], Juliette Lengellé[5], Jean-Christophe Baurens[1], Olivier Garsmeur[1], Alberto Cenci[2], Bertrand Pitollat[1], Angélique D'Hont[1], Manuel Ruiz[1], Mathieu Rouard[2], Stéphanie Sidibe-Bocs[1]

[1] CIRAD, UMR AGAP, France
[2] Bioversity International, France
[3] Montpellier SupAgro, UMR 1334, France
[4] IRD, France
[5] Inra, France

Presenter: Gaëtan Droc

Banana is not only one of the world's favorite fruits but also one of the most important crops for developing countries. The banana reference genome sequence (Musa acuminata) was released recently. Given its taxonomical position, the sequence has particular comparative value highlighted in the recent genome publication providing new insights regarding the evolution of monocotyledons. This study has been enhanced by a number of tools and resources described in this publication. Here, we present all the tools that are available to harness the reference genome sequence and to support post-genomic efforts, such as transcriptomic and metabolomic pathways studies. Several uses cases illustrate how the banana genome hub can be used to study gene families. Overall, with this collaborative effort we discuss the importance of the interoperability towards data integration between existing information systems. Database URL: http://banana-genome.cirad.fr/

76

IntAct Database in 2013

Margaret Duesbury, Team IntAct

EMBL-EBI, United Kingdom

Presenter: Margaret Duesbury

IntAct is an open-source, open data molecular interaction database populated by data either curated from the literature or from direct data depositions. Two levels of curation are now available within the database, with both IMEx-level annotation and less detailed MIMIx-compatible entries are currently supported. As from January 2013, IntAct contains approximately 300,000 curated binary interaction evidences from over 6000 publications, and includes 12,000 interacting domain annotations. IntAct is an active contributor to the IMEx consortium (http://www.imexconsortium.org). Records curated by the MPIDB Database, which ceased active curation in 2012, have now been transferred to IntAct. IntAct source code and data are freely available at http://www.ebi.ac.uk/intact . Intact exports filtered, high confidence protein interaction data to other databases, including UniProt, NextProt and GOA, and are happy to extend this service to other databases on request.

77

New views of Rfam

Ruth Eberhardt, Sarah Burge, Jennifer Daub, John Tate, Alex Bateman

EMBL-EBI, United Kingdom

Presenter: Ruth Eberhardt

Rfam is a database of non-coding RNA families, each represented by a covariance model (CM) and a multiple sequence alignment. Release 11.0 of Rfam (August 2012) contains 2208 families annotating over 6 million sequence regions. This represents a significant growth from release 10.1 which contained 1973 families and annotated almost 2.8 million regions in the rfamseq database. Rfam 11.0 contains several new features. We have introduced a sunburst representation of the sequence distribution of each family. This view enables users to easily select a taxonomic subset of sequences and either download or align them to the CM. We have added a Biomart to assist users who wish to run more complex queries than those available through our website. Another new feature is a RESTful interface, to allow easier programmatic interaction with the Rfam database. More recently, we have added R-chie arc diagrams to our suite of secondary structure representations. The challenge presented by the ever-increasing size of our underlying sequence database has necessitated changes in the types of alignment that we supply for our largest families. Smaller genome-based alignments are now provided for these families, and in future we will provide these for all families.

78

Enhancing ChEBI, the reference database and ontology for biologically relevant chemistry

Marcus Ennis, Janna Hastings, Paula de Matos, Adriano Dekker, Bhavana Harsha, Namrata Kale, Venkatesh Muthukrishnan, Gareth Owen, Steve Turner, Mark Williams, Christoph Steinbeck

EMBL-EBI, United Kingdom

Presenter: Marcus Ennis

ChEBI (Ch emical E ntities of B iological I nterest; www.ebi.ac.uk/chebi) [1] is a database and ontology for low-molecular-weight chemical entities relevant for understanding and intervening in biological functioning. ChEBI continues to grow in annotated content, with the February 2013 release containing over 31,000 fully annotated entities. Throughout 2012 we have seen a steady increase in the number of submissions created directly in the database by our users, with currently more than 2800 of the current total of annotated entries having been submitted via this route. A large-scale ongoing curation effort in collaboration with the La Jolla Institute for Allergy and Immunology focuses on annotating compounds relevant for immunology, while, together with partner database Metabolights [2], a further major curatorial effort is currently being directed towards the annotation of a set of >4000 natural products using structural and species information extracted from the literature. The harnessing of information on the biological roles of drugs, the addition of citations (both general and species-specific), the incorporation of text definitions which differentiate a compound from its immediate parents in the taxonomic hierarchy, and the addition of database links to several new resources including MetaCyc, ChemSpider and Wikipedia all represent further advances in manual curation initiated during 2012. Improvements to the ontology incorporated during the past year include its alignment with the Open Biomedical Ontologies (OBO) Foundry-recommended upper level Basic Formal Ontology (BFO), alignment with the classification of chemical-involving processes in the Gene Ontology (GO), and a dynamic new interactive graph-based ontology visualisation. References 1. Hastings, J. et al ., Nucl. Acids Res. (2013) 41 (D1): D456-D463 (doi:10.1093/nar/gks1146). 2. Haug, K. et al ., Nucl. Acids Res. (2013) 41 (D1): D781-D786 (doi:10.1093/nar/gks1004).

79

Manual curation of variation data and genetic diseases in UniProtKB/Swiss-Prot

Maria Livia Famiglietti[1], EBI UniProt consortium[2], PIR UniProt consortium[3]

[1] SIB Swiss Institute of Bioinformatics, Switzerland
[2] EMBL-EBI, United Kingdom
[3] Protein Information Resource, United Kingdom

Presenter: Maria Livia Famiglietti

UniProtKB/Swiss-Prot, the manually curated section of the UniProt Knowledgebase (http://www.uniprot.org/) records information on human protein variants and genetic diseases. Relevant data are manually retrieved from publications, particularly focusing on characterized single amino-acid polymorphisms (SAPs) and their functional consequences and association with disease. The complete index of all SAPs and their classification is available at http://www.uniprot.org/docs/humsavar. Release 2013_01 contains 67’262 SAPs, classified into disease–associated variants, variants of unknown pathological significance, and benign polymorphisms. In total, 6% of all UniProt variants are associated with annotations describing their impact on protein function. In order to facilitate the integration of our curated variant data with that from other resources, UniProt SAPs are mapped to reference nucleotide sequences from RefSeq and Locus Reference Genomic (LRG) sequences (http://www.lrg-sequence.org/) and are submitted to specialized Locus Specific Databases (LSDBs) as well as to the dbSNP repository. This work is carried out in the frame of Gen2Phen (http://www.gen2phen.org/), a collaborative project aiming to unify genetic variation databases. Curated information on variants is linked to disease descriptions in UniProtKB/Swiss-Prot records (annotated in the “Involvement in disease” subsection of the “General Annotation” section). We are currently preparing an index of genetic diseases containing disease names, synonyms, descriptions, and cross-references to OMIM phenotypes and MeSH terms. We also plan to standardize the annotation of functional consequences of variants using selected terms from existing ontologies.

80

Programmatic Curation of BioSamples Database

Adam Faulconbridge, Tony Burdett, Helen Parkinson, Marco Mrandizi, Mikhail Gostev, Rui Pereira, Ugis Sarkans, Alvis Brazma

EMBL-EBI, United Kingdom

Presenter: Adam Faulconbridge

The BioSamples Database at EBI is a database of sample meta-information used within EBI (e.g. ArrayExpress, ENA, PRIDE). It contains over 12,000,000 samples in over 500,000 groups from 10 different sources and is expected to continue to grow in size. 30,000 samples are classed as "reference" - highly curated samples reused elsewhere e.g. cell lines, mouse strains. Problems addressed by BioSamples Database include making submissions to EBI databases easier by referencing existing samples rather than manually re-entering meta-information, and assisting the identification of appropriate samples for meta-analyses. Each sample is described by a set of attributes that are composed of a key-value pair as well as optional controlled vocabulary / ontology mapping and units. Attributes are also used to specify relationships between samples. Multiple sample records corresponding to a single physical sample are connected by "same as" relationships e.g. RNA-seq in both ArrayExpress and ENA. Other relationships are "child of" for family groups and and “derived from” for temporal relationships. Relationships are used by programmatic curation to infer information and prevent contradictory data being stored. This has identified samples in source resources that need further manual curation in order to solve inconsistencies. Sample attributes are harmonized over the different sources and mapped to ontologies (e.g. EFO) where possible. This process is automated due to the large number of samples as well as the flow of new and updated samples in the sources. Typical examples of such harmonization include species name (e.g. replacing "human" with "Homo sapiens"), case differences (e.g. "Strain" vs "strain") as well as spacing and / or punctuation (e.g. "wildtype" , "wild-type" and "wild type"). As part of the attribute harmonization process, new relationships may be inferred such as where assays have used a reference layer sample.

81

Literature curation at dictyBase

Petra Fey, Robert Dodson, Kerry Sheppard, Rex Chisholm

Northwestern University, United Kingdom

Presenter: Petra Fey

dictyBase is the model organism database for the social amoeba Dictyostelium discoideum . As of January 2013, the dictyBase literature corpus is 7,323 of which nearly 1,900 have been annotated. While these numbers are small compared to larger research fields, we only have 2 curators whose responsibilities are manifold: User requests, HTML page updates, grant and paper writing, supervising Dicty Stock Center (DSC) operations, literature, and gene curation. In order to keep up with new literature we needed to make curation more efficient. From literature we primarily annotate gene names, products, strains, phenotypes, and Gene Ontology (GO) terms. Strain annotations are of importance at dictyBase because we also host the DSC, a repository for Dictyostelium strains and plasmids. Strain annotation is often done in conjunction with the DSC staff who request newly published strains for deposit and are annotated with several controlled vocabularies, such as strain descriptors and mutagenesis types. Phenotypes are then attached to the strains, which in turn are linked to genes. We are developing our own phenotype ontology, a pre-composed ontology from Dictyostelium anatomy terms, GO biological processes, and quality terms. Gene ontology is annotated using the Protein2GO tool provided by the EBI. In collaboration with Wormbase we began to integrate a textpresso pipeline to assist cellular component GO annotations. Strain curation is very time consuming because information is often not well represented in papers. Therefore we recently initiated a community curation trial, in which researchers annotate their newly published papers prior to curator review. This has been very helpful particularly for strain and phenotype curation. As a small database in which each person has many obligations, literature curation is becoming more dependent on collaborators and efficient tools. dictyBase is funded by NIH GM64426, GM087371 and HG0022.

82

Biases in the Experimental Annotations of Protein Function and their Effect on Our Understanding of Protein Function Space

Alexandra Schnoes[1], David Ream[2], Alexander Thorman[2], Patricia Babbitt[1], Iddo Friedberg[2]

[1] University of California, San Francisco, United States of America
[2] Miami University, United States of America

Presenter: Iddo Friedberg

The ongoing functional annotation of proteins relies upon the work of curators to capture experimental findings from scientific literature and apply them to protein sequence and structure data. However, with the increasing use of high-throughput experimental assays, a small number of experimental studies dominate the functional protein annotations collected in databases. Here we investigate just how prevalent is the ``few articles -- many proteins'' phenomenon. We examine the annotation of UniProtKB by the Gene Ontology Annotation project, and show that the distribution of proteins per published study is exponential, with 0.14% of articles annotating 25% of the proteins in UniProt-GOA. Since each of the dominant articles describes the use of an assay that can find only one function or a small group of functions, this leads to substantial biases in what we know about the function of many proteins. Mass-spectrometry, microscopy and RNAi experiments dominate high throughput experiments. Consequently, the functional information derived from these experiments is mostly of the subcellular location of proteins, and for participation in embryonic developmental pathways. For some organisms, the information provided by different studies overlap by a large amount. We also show that the information provided by high throughput experiments is less specific than those provided by low throughput experiments. Given the experimental techniques available, certain biases in protein function annotation due to high-throughput experiments are unavoidable. Knowing that these biases exist and understanding their characteristics and extent is important for database curators, developers of function annotation programs, and anyone who uses protein function annotation data to plan experiments.

83

Dog gene annotation in Ensembl

Carlos García Girón[1], Daniel Barrell[1], Andreas K. Kähäri[1], Thibaut Hourlier[1], Fergal Martin[1], Rishi Nag[1], Simon White[2], Amonida Zadissa[3], Bronwen Aken[1], Steve Searle[1]

[1] Wellcome Trust Sanger Institute, United Kingdom
[2] Baylor College of Medicine, United Kingdom
[3] EMBL-EBI, United Kingdom

Presenter: Carlos García Girón

Ensembl (www.ensembl.org) provides automatic genome annotation for over 50 vertebrate species including dog. New multiple species alignments, gene trees, regulatory and variation features are made freely available online in each release. This data can be accessed via the Ensembl Browser, MySQL databases, a Perl API and the BioMart query system. Canis lupus familiaris (dog) is a model organism which shares with human a wide range of diseases such as cancer, diabetes and obsessive-compulsive disorder. Hence the impact of its gene annotation on comparative studies is important for understanding diseases common in both species. The latest dog assembly, CanFam3.1, was produced in September 2011 by the Broad Institute of MIT and Harvard. The gene annotation was made public in Ensembl release 68 (July 2012). The 2.4Gb high-coverage dog genome was annotated using the Ensembl genebuild pipeline, incorporating RNA-seq data from 10 different tissues provided by the Broad Institute. The initial set of analyses masked 44.7% of the genome as repeat features. Subsequent analyses on the masked genome included CpG islands and transcript start sites identification together with ab initio gene predictions and alignment of protein, cDNA and EST sequences from UniProt, RefSeq and ENA. This led to the construction of a gene set following a process where various levels of evidence, parameters and filtering steps were taken into account. Evidence from dog proteins, cDNA and RNA-seq data together with mammalian proteins contributed to the final gene set of 23,630 genes. Of these, 19,856 are coding genes, 3,794 are short non-coding RNA genes and 950 pseudogenes. 8,968 of the genes were exclusively annotated using the RNA-seq data. Full annotation of the dog genome is available at www.ensembl.org, along with comparative genomic, regulation and variation resources.

84

Proteomic data integration pipeline in neXtProt

Isabelle Cusin, Guilaine Argoud-Puy, Monique Zahn, Alain Gateau, Anne Gleizes, Ying Zhang, Lydie Lane, Amos Bairoch, Pascale Gaudet

SIB Swiss Institute of Bioinformatics, Switzerland

Presenter: Isabelle Cusin, Pascale Gaudet

A major current challenge in biological sciences is the accessibility of data, given the large volume of data produced by high-throughput techniques. Most large proteomics data sets are published in supplementary materials that are difficult to exploit. Moreover, different experiments are difficult to compare, because they are done either in different conditions, or using different instruments, or analyzed using different methods. Finally, not all data is of equal quality; although some low-confidence data may contain 'gold nuggets', it is important to have a way to access high quality data in a centralized resource. neXtProt (www.nextprot.org/) is a protein knowledge platform built upon the corpus of human data from Swiss-Prot. In addition, it contains data originating from a variety of high-throughput approaches such as microarray, antibodies screens, proteomics, and interactomics. One main focus in neXtProt is the integration of protein and PTM site identifications by mass spectrometry. The data is obtained from large-scale studies via direct submissions or original papers, carefully selected on a quality basis. For peptides, we have used an additional source of data, PeptideAtlas (http://peptideatlas.org/). For each data set, data points are attributed a confidence level as follows: gold-quality data has a very high level of confidence, typically < 1% error rate; silver data is of high confidence, i.e. < 5% error rate. Any data with lower confidence is assigned "bronze" quality and not integrated in the database. Each annotated peptide and PTM site is accompanied by metadata that provides detail on the experimental setup using controlled vocabularies. In the past two years, we have integrated nearly 20,000 PTM sites (N-glycosylation, phosphorylation, S-nitrosylation, ubiquitination and sumoylation), which corresponds to close to 8,000 new PTM annotations. We have also mapped 261,000 identified peptides to 13,375 proteins.

85

Submissions at the European Nucleotide Archive

Richard Gibson, European Nucleotide Archive

EMBL-EBI, United Kingdom

Presenter: Richard Gibson

The European Nucleotide Archive (ENA; http://www.ebi.ac.uk/ena) is the primary European resource for capturing and presenting nucleic acid sequence data to the scientific community. Public data is exchanged within the International Nucleotide Sequence Database Collaboration (INSDC). The core data within the ENA comprises of next generation sequencing (NGS) reads and traditional capillary sequences. Although raw reads dominate in data volume, assembled and annotated sequences dominate in terms of submitter and user numbers. Despite the core divisions, ENA is working to unify the system used for the submission of nucleotide data. Pipeline and programmatic services exist for regular large-scale submitters whereas medium- to small-scale submitters can use the online submission service, Webin. This system is composed of two concepts: a manual tool and a pre-tailored checklist system. The manual tool is the traditional Webin that allows submitters to describe their assembled sequences fully in the controlled language and syntax of the INSDC. The checklist system has been designed by curators for the most common sequence annotation types encountered at EMBL-Bank; it also serves as a simple tool for providing standards-compliant NGS metadata. All data, regardless of route into the ENA, undergoes a degree of automatic validation. This process feeds back both accidental errors and overlooked issues with the submitted data, which the user must correct before being able to complete a submission. The ENA has a small team of curators who are continuously adding to the validator's biological rule-base, thus increasing overall database consistency and reducing time spent on repetitive tasks. Curator effort can then be focussed on applying their trained knowledge to interrogate and edit biological data, to provide outreach and helpdesking, and to participate in the development of standards and best practices in capturing, storing and presenting sequence metadata.

86

The ISA framework and its user community: curate, analyse, share, publish and visualize biodata

Alejandra Gonzalez-Beltran, Eamonn Maguire, Philippe Rocca-Serra, Susanna-Assunta Sansone

Oxford e-Research Centre, University of Oxford, United Kingdom

Presenter: Alejandra Gonzalez-Beltran, Eamonn Maguire, Philippe Rocca-Serra, Susanna-Assunta

Many community-based initiatives published minimum reporting guidelines, terminologies and formats (generally referred to as 'standards')[1] to structure and curate datasets, enabling data annotation to varying degrees. Funding agencies and publishers embrace the concept that standards are pivotal to enriching the annotation of the relevant entities (genes, metabolites) and the experimental steps (e.g. provenance of study materials, assay types), to ensure that shared investigations are comprehensible, reproducible and reusable in principle. But how can we enable bioscience researchers and curators to make use of existing community standards The Investigation/Study/Assay (ISA)[2,3] open source, metadata-tracking framework is a successful example developed and supported by the growing ISA Commons community[4]. The ISA framework includes both a general-purpose file format and a software suite to tackle the harmonization of the structure of bioscience experimental metadata (e.g., provenance of study materials, technology and measurement types, sample-to-data relationships) in an increasingly diverse set of life science domains (including metabolomics, (meta)genomics, proteomics, system biology, environmental health, environmental genomics and stem cell discovery) and also enables compliance with the community standards. The ISA Commons illustrates how the synergy between research and service groups in academia[5,6], in industry (e.g. at The Novartis Institutes for BioMedical Research and at Janssen Pharmaceuticals, a company of Johnson & Johnson) and in the publishing domain (GigaDB for GigaScience [7]), is pivotal to build a network of data collection, curation, and sharing solutions that progressively enable the 'invisible use' of standards.

1. www.biosharing.org
2. Bioinformatics 15;26(18):2354-6 (2010)
3. Bioinformatics (2013)
4. Nat Genet 27 44(2):121-126 (2012)
5. Nucleic Acids Res 40:D984-91 (2012)
6. Nucleic Acids Res 41:D781-6 (2013)
7. www.gigasciencejournal.com

87

On the reproducibility of science

Nicole Vasilevsky, Matt Brush, Melissa Haendel

OHSU, United States of America

Presenter: Melissa Haendel

The scholarly communication cycle is the process where scholars create, share, preserve, and extend their research results. This cycle is the cornerstone of biomedical research: the literature is where we put our collective knowledge about science. Despite improvements in document accessibility, a large problem still exists within the context of this literature - authors simply do not uniquely reference research resources enough to enable adequate research reproducibility. Severe faults in experimental reproducibility were highlighted in a Nature paper by researchers at Amgen, who found that only 11% of the academic research in the literature was reproducible by their groups (Begley and Ellis 2012). In part, a lack of unique reference to research resources, such as antibodies and model organisms, makes it difficult or impossible to reproduce science. In order to better understand the magnitude of this problem, we designed an experiment to evaluate the "identifiability" of research resources in the biomedical literature. We evaluated recent journal articles in the fields of Neuroscience, Developmental Biology, Immunology, Cell and Molecular Biology and general Biology, selected randomly based on a diversity of journal impact factors, publisher, and resource reporting guidelines. We attempted to uniquely identify model organisms (mouse, rat, zebrafish, worm, fly and yeast), antibodies, knockdown reagents (morpholinos or RNAi), DNA constructs and cell lines. This effort was informed by a well-defined a set of attributes describing the provenance, experimental features, and associated identifiers for each resource type. For example, antibody metrics included documentation of its source organism, immunogenic target and any catalog or registry identifier linked to the reagent. We will present the results of this experiment and our recommendations to journal editors, publishers, reviewers and authors on how to best approach this significant problem in science today.

88

MONARCH: A new discovery system for integration and exploration of genotype-phenotype resources

Melissa Haendel[2], Anita Bandrowski[1], Matt Brush[2], Jeff Grethe[1], Amarnath Gupta[1], Harry Hochheiser[3], Maryann Martone[1], Carlo Torniai[2], Nicole Vasilevsky[2], Nicole Washington[4], Chris Mungall[4]

[1] University of California San Diego, United States of America
[2] Oregon Health & Science University, United States of America
[3] University of Pittsburgh, United States of America
[4] Lawrence Berkeley National lab, United States of America

Presenter: Melissa Haendel

Model systems are the cornerstone of biomedical research that progresses our understanding of biological and disease processes. There are many databases that contain unparalleled content about genetics, genomics, anatomy, environmental interactions, phenotypes, and developmental processes of model systems. However, sifting through these vast stores of biomedical data can be a daunting task for researchers seeking to improve understanding of human disease mechanisms towards therapeutic discovery. To address this challenge, we are building a novel discovery system that will include information about model organisms and in vitro models, tools that exploit similarities between phenotypes, organisms, and human diseases, and an innovative user interface to visually explore and filter on the semantic relationships between models and disease that leverage links between genes, pathways, gene expression, protein and genetic interactions, orthology, assays and publications. Building off of the Neuroscience Information Framework (NIF) DISCO platform, our ingest pipeline imports data from multiple genetic and phenotype resources, seeded initially with data from the OMIM, MGI, and ZFIN databases. To integrate genotype-phenotype associations derived from varied resources, each of which has their own data model, we first developed an integrated genotype model in OWL. This GENO ontology models genotype and its components (e.g. alleles, SNPs, genes, background strains, etc.). Each genotype-phenotype resource is curated to GENO using the NIF Concept Mapper, phenotypes are bridged between model organism and human anatomy with Uberon, and the Sequence Ontology is used for genomic annotation interoperability, which together lays the groundwork for phenotypic similarity comparison. The resulting semantically mapped data will be made publicly accessible via the user interface, services, and a SPARQL endpoint to support maximal exchange and interoperability with contributing sources.

89

Community Annotation of Genomes using Zmap

Jane Loveland[1], Jennifer Harrow[2], Matthew Hardy[1]

[1] HAVANA, WTSI, United Kingdom
[2] Wellcome Trust Sanger Institute, United Kingdom

Presenter: Matthew Hardy

Human and Vertebrate Annotation and Analysis (HAVANA) group, Wellcome Trust Sanger Institute (WTSI), Hinxton, Cambridgeshire CB10 1HH, UK Manual annotation plays an important role in providing “added value” to the genomes of non-model organisms. Since manual annotation is an expensive resource and funding is limited we help communities to use our in-house tools and guidelines to produce high quality annotation that can be merged into an Ensembl gene build. Our annotation tools (Otterlace/Zmap) can be used remotely by external collaborators and are available for Mac and Linux. We provide research communities with software and training. The data is either QC’d internally by experienced annotators (i.e. the Gatekeeper approach) or users are allowed to deposit their data directly when work has been approved of sufficient standard (e.g. Blessed Annotator).(Loveland et al. 2012). We have annotated over 1500 pig genes on unfinished genomic contigs as part of a community annotation group of over 30 annotators worldwide, forming the Immune Response Annotation Group (IRAG). The VEGA website has been updated to include the new pig build (10.2) including the IRAG annotation which has been merged into the Ensembl geneset. Following the success of this model we are now working with the Rat community including RGDB, to update prioritized regions of the new rat genome assembly. To launch this endeavor, we have organized a community annotation meeting in March amongst the rat genome researchers to introduce them to our tools and discuss where how to prioritse regions to be manually annotated, and will feedback on its success.

Reference: Jane E. Loveland*, James G.R. Gilbert, Ed Griffiths and Jennifer L. Harrow* (2012). Community gene anotation in practice. Database, Vol. 2012, Article ID bas009

90

Phenotype Curation in PomBase

Midori Harris[1], Antonia Lock[2], Jürg Bähler[2], Stephen Oliver[1], Valerie Wood[1]

[1] University of Cambridge, United Kingdom
[2] University College London, United Kingdom

Presenter: Midori Harris

PomBase, the recently established online fission yeast resource, has made the comprehensive and detailed representation of phenotypes a high priority. To support highly specific phenotype annotation, we are actively developing the Fission Yeast Phenotype Ontology (FYPO), a modular ontology that uses several OBO Foundry as building blocks, including the phenotypic quality ontology PATO, the Gene Ontology (GO), and Chemical Entities of Biological Interest (ChEBI). Phenotype curation is featured in Canto, the PomBase community curation tool. Canto and the Chado database underlying PomBase support annotation of specific alleles, singly or in combinations (e.g. double or triple mutants), to FYPO terms, along with supporting evidence. Relevant experimental conditions can be captured, and targets, expressivity and penetrance can be represented by annotation extensions. Phenotype annotations are displayed on PomBase gene pages. FYPO terms can be searched by name or ID in the PomBase advanced search. Over 6000 legacy annotations inherited from GeneDB have been converted to FYPO terms, and new annotations are being made in Canto by both PomBase curators and fission yeast community researchers, with over 2200 created to date.

91

Reactome: a human pathway knowledgebase

Robin Haw[1], Reactome Team[2], Henning Hermjakob[3], Peter D'Eustachio[4], Lincoln Stein[1]

[1] OICR, Canada
[2] Reactome, United Kingdom
[3] EMBL-EBI, United Kingdom
[4] NYUMC, United States of America

Presenter: Robin Haw

Reactome is an open-source, free access, manually curated and peer-reviewed biological pathway knowledgebase. Information is authored by expert biological researchers, maintained by the Reactome editorial staff, and cross-referenced to publicly available web-based informatics resources. The Reactome data model describes life processes ranging from metabolism to cell signaling, in a single internally consistent, computationally navigable format. Recent extensions of our data model accommodate the annotation of cancer and other disease processes. To support the graphical representation of disease pathways, we have adapted our pathway browser to display disease variants and events in a way that allows comparison with wild type pathway, and shows connections between disease and other pathways. The pathway diagrams can be overlaid with protein and chemical interactors or expression measures from external sources or submitted datasets. Reactome includes tools for pathway enrichment analysis and large-scale data querying. Pathway data can be exported in several formats including SBML, BioPAX, SBGN, and derived interactions. Reactome content, the database and software interface are freely available. See www.reactome.org for more information.

92

Increased curation efficiency is utilized to provide expanded pathway information at the updated Rat Genome Database Pathway Portal

G. Thomas Hayman, Pushkala Jayaraman, Victoria Petri, Marek Tutaj, Weisong Liu, Jeff De Pons, Melinda Dwinell, Mary Shimoyama

Medical College of Wisconsin, United States of America

Presenter: G. Thomas Hayman

The Rat Genome Database Pathway Portal provides pathway information that is easily accessed, visualized and integrated with other data types. This includes annotations for human, rat and mouse genes, pathway information imported by pipelines from other databases using term synonyms and interactive diagrams along with pathway suites and suite networks; all are interconnected via the pathway ontology (PW). The PW allows for standardized annotation of genes to any pathway type. It also allows navigation between gene and ontology report pages and between connected or related pathways in diagram pages and suites. The latter feature allows users to journey across the pathway landscape. Diagram pages for pathway terms present the diagram with a legend for objects and relationships and a link to the term's ontology report page. Objects within the diagram are linked to gene report pages for genes, to lists of genes for members of a class, to PubChem or ChEBI for small molecules or to the ontology report page for other pathway(s) shown in the diagram. An expandable description is provided that contains Pfam entries for protein domains and links to KEGG, Reactome and GO term entries. Relevant PubMed reference links are provided as well as the mapped path to the term in the ontology tree. A new, dual-functionality web application has been developed that composes the diagram page elements. Curators input the description and the diagram, references and additional pathway objects. The application then parses the database to generate tables of rat, human and mouse pathway genes which include genetic information, and analysis tool and reference links. Additional tables generated by the application house disease, phenotype and other pathway annotations to pathway genes which can be toggled between listing alphabetically by gene or by disease, pathway or phenotype. The application efficiently increases the information content of diagram pages while expediting the publication process.

93

JBrowse: a next-generation genome browser.

Gregg Helt

Lawrence Berkeley National Laboratory, United States of America

Presenter: Gregg Helt

JBrowse is the next generation web-based genome browser from the GMOD project built from the ground up with JavaScript and HTML5. Like its predecessor GBrowse, it is designed to be a highly customizable viewer for presenting and integrating genomic data dynamically at any level of detail from base pairs to chromosomes. JBrowse offers advanced integration, fluid interactivity, and native support for very large datasets (such as deep-coverage resequencing alignments) that were once possible only for desktop-based genome viewers while at the same time enabling simple, reproducible deployments and seamless data interchange made possible by the cloud. For large installations with thousands of available data tracks, JBrowse incorporates a track selector with faceted browsing, so that previously rich-but-overwhelming lists of tracks (e.g. spanning many experimental protocols, computational analyses, cell lines, and collaborative groups) become searchable in real time. In the talk we will demonstrate JBrowse's capabilities in more detail and outline future development plans. JBrowse is open source, free for academic and commercial use, and paid enterprise support is available.

94

An Ontology for formal representation of Drug Drug Interaction Knowledge

Maria Herrero Zazo, Isabel Segura-Bedmar, Paloma Martinez Fernandez

University Carlos III Madrid, Spain

Presenter: Maria Herrero Zazo

Ontologies are useful tools in text mining research tasks as a source of specialized vocabulary of terms and relationships in a given domain. Furthermore, formal knowledge representation provided by ontologies can be applied for new knowledge inference, which can be exploited for biomedical research purposes. Drug-drug interactions (DDIs) are common adverse drug reactions having an important impact on patient safety and healthcare costs. To the best of our knowledge, there are only two ontologies on drug interactions: DIO (Drug Interaction Ontology)is a formal representation of drug pharmacological actions, depicted by drug-biomolecules interactions, and the PK ontology, which was developed for the representation of drug pharmacokinetic related information, including an important type of DDIs: those occurring through a pharmacokinetic mechanism. However, DDIs are complex phenomenon related with different types of information other than pharmacokinetics. We propose the development of a comprehensive drug interaction ontology that will provide a formal representation of the knowledge related to drug interactions. It will include those entities involved in a drug interaction (drugs, metabolites, proteins) and those possible relations among them, including information about their mechanism, their molecular and clinical effects, the advice or management related information, etc. The first step in the framework of this ongoing process arose from the development of a manually annotated corpus for the DDI Extraction 2013 task, whose main goal is to provide a common framework for evaluation of information extraction techniques applied to the detection of drug-drug interactions from biomedical texts. Initial stages in ontology creation process were carried out through the study of the documents within the corpus and the creation of an annotation schema, which led to the identification of basic concepts in the DDI domain and to a prior analysis of relations between them.

95

The Zebrafish Anatomy Ontology (ZFA) and its Usage in Curating Phenotype Data at ZFIN

Yvonne Bradford, Doug Howe, Christian Pich, Leyla Ruzicka, Brock Sprunger, Ceri Van Slyke

Zebrafish Model Organism Database, United States of America

Presenter: Christian Pich

ZFIN produces the Zebrafish Anatomy Ontology (ZFA) and uses it in conjunction with a variety of ontologies to facilitate curation and data searches. ZFIN will begin adding Gene Ontology (GO) logical definitions to the ZFA utilizing logical definition files generated by GO. This approach increases the accuracy of phenotype searches where the curated phenotype uses a GO process (e.g. retina development) but the query uses a ZFA structure (e.g. eye) related to the GO process. Here we describe the current status of the ZFA ontology, the supporting database schema, and how phenotype data is annotated using entity:quality (EQ) syntax and entity post-composition.

96

Faceted Search Strategies at ZFIN

Doug Howe, David Fashena, Kevin Schaper, Leyla Ruzicka, Christian Pich

Zebrafish Model Organism Database, United States of America

Presenter: Doug Howe

Most biological databases, including ZFIN, provide parametric search forms. In a parametric search, the user fills out one or more entry fields to constrain search results with respect to specific parameters. One might search for genes whose name contains "hedgehog", or an antibody that labels the kidney. This approach works well enough for data sets with only a few searchable parameters, or in which typical searches produce only a small number of search results. However, modern biological data has grown in volume and complexity, revealing the limitations of traditional parametric search. A) Too many results are returned for a user to reasonably sort and process. B) As more parameters are required to specify a search, it becomes increasingly likely the search will return no results. C) Parametric searches offer no way to explore or navigate results, providing no window into the data to inform the next search users may want to try. To address these issues, faceted searching has become common on the web - found in a wide variety of shopping sites, library sites etc. Faceted search provides immediate feedback to users about how much data is associated with each new query they may run. This allows users to explore the data without getting empty result sets. Here we describe our efforts to produce a faceted search of ZFIN data, backed by the open-source search server Solr. We discuss boosting strategies for sorting results, UI design choices and the special challenges posed by the incorporation of ontology terms as facets. We believe that faceted search will provide a powerful new way to explore the ZFIN database.

97

Protein2GO: A curation tool for Gene Ontology

Rachael Huntley, Tony Sawford, Prudence Mutowo-Meullenet, Yasmin Alam-Faruque, Emily Dimmer, Maria-Jesus Martin, Claire O'Donovan, Rolf Apweiler

EMBL-EBI, United Kingdom

Presenter: Rachael Huntley

The UniProt-Gene Ontology Annotation (UniProt-GOA) project is a public resource that provides high-quality Gene Ontology (GO) annotations to proteins in the UniProt Knowledgebase (UniProtKB). The assignment of GO terms to UniProtKB records is an integral part of UniProt biocuration. Protein2GO has been the biocuration tool for GO annotation at the EBI for over 10 years. Improvements have been made during this time to cater to the needs of biocurators as well as to keep up to date with evolving GO annotation guidelines. Protein2GO is now in a mature state which allows biocurators to easily and quickly capture the data integral to a GO annotation as well as optional extra information that the GO Consortium has recently introduced into the annotation format. The tool is optimised for assigning GO terms to UniProtKB accessions, however, due to the recent adoption by the GO Consortium of Protein2GO as a common biocuration tool for protein functional annotation, it has been adapted to allow biocurators from model organism databases to search for protein entries using commonly used identifiers, gene names or synonyms specific for their species or database. Protein2GO incorporates many quality control checks to prevent inappropriate annotation statements being added to the database. We have recently incorporated an added-value functionality that suggests additional or more specific GO terms for the protein being curated. In addition, we have used existing annotations and additional data in UniProtKB entries to alert biocurators to inappropriate annotations. Another new development enables a biocurator to dispute any of the manually generated annotations in the database. Protein2GO is a robust and extendable biocuration tool that has proven invaluable for the assignment of GO functional data to UniProtKB entries. At present over 60 biocurators use Protein2GO and this is likely to increase in the coming years as we provide access to further GO Consortium members.

98

PCorral – interactive mining of protein interactions from MEDLINE

Antonio Jimeno-Yepes[1], Chen Li[2], Miguel Arregui[2], Harald Kirsch[2], Dietrich Rebholz-Schuhmann[3]

[1] National ICT Australia, Australia
[2] EMBL-EBI, United Kingdom
[3] EMBL-EBI / University of Zürich, United Kingdom

Presenter: Chen Li

The extraction of information from the scientific literature is a complex task – for researchers doing manual curation as well as for automatic text processing solutions. The identification of protein-protein interactions (PPIs) is such a task that requires the extraction of protein named entities and their relations. Semi-automatic interactive support is one approach to combine both solutions for efficient working processes to generate reliable database content. In principle, the extraction of PPIs can be achieved with different methods that can be even combined to deliver data of high precision and high recall in different combinations at the same time. Interactive use can be achieved, if the analytical methods are fast enough to process the retrieved documents. PCorral allows interactive mining of PPIs from the scientific literature helping curators to cope with the sustained growth of MEDLINE. Advantages of PCorral are both its versatile and interactive use to generate either high recall or high precision results for the curation work. The underlying components of PCorral process the documents on-the-fly. The human interface summarizes the identified PPI results and the involved entities are linked back to relevant resources and databases. Finally, the core PPI modules are available as a web service from Whatizit.

99

Collaborative curation for Identifiers.org

Camille Laibe[1], Nicolas Le Novère[2], Henning Hermjakob[1], Nick Juty[1]

[1] EMBL-EBI, United Kingdom
[2] Babraham Institute, United Kingdom

Presenter: Camille Laibe

The MIRIAM Registry (http://www.ebi.ac.uk/miriam/) records information about collections of data in the life sciences, as well as where it can be obtained. This information is used, in combination with the resolving infrastructure of Identifiers.org (http://identifiers.org/), to generate globally unique identifiers, in the form of Uniform Resource Identifier (URIs). These identifiers are now widely used to provide perennial cross-references and annotations. The growing demand for these identifiers results in a significant increase in curational efforts to maintain the underlying registry. This requires the design and implementation of an economically viable and sustainable solution able to cope with such expansion. We briefly describe the Registry, the current curation duties entailed, and our plans to extend and distribute this workload through collaborative and community efforts.

100

Linking the Literature to Biomolecular Databases: A comprehensive analysis based on database accession numbers in full text

Senay Kafkas, Jee-Hyub Kim, Johanna McEntyre

EMBL-EBI, United Kingdom

Presenter: Senay Kafkas

Biomolecular databases are a core component of life science research. Articles cited in those databases convey useful functional information about those database records. In this study, we present a comprehensive analysis on data citation practice based on explicit database accession numbers in the Open Access subset of Europe PubMed Central (OA-ePMC) full text articles (http://europepmc.org/). By focusing on three major databases, the European Nucleotide Archive (ENA), UniProt and Protein Data Bank (Europe) (PDBe), we explore the extent to which accession numbers are annotated by publishers and to what extent text mining can further identify accession numbers in the text narrative. Furthermore, we compare these findings to the citation of articles from databases. To the best of our knowledge, this is the first comprehensive study on accession number citation analysis in the OA-ePMC full text articles. Results reveal that our text mining approach substantially improves the number of accession numbers discovered in articles. It suggests thousands of new links from the literature to databases that are not in the set of links gathered from databases to the literature. Assessment of our text-mining pipeline shows that it achieves very high performance (precision and recall values are above 95%). An extended version of the text-mining pipeline covering other databases, such as Ensembl, ArrayExpress and Pfam, which are also widely cited in the literature, is being integrated into the Europe PMC infrastructure. This new facility of Europe PMC will bring more opportunity to further integrate research activities in life sciences and provide the basis of the development of new tools that may assist in curation processes.

101

A Web Service for Accession Number Mining in Full-Text Articles for Biocuration

Jee-Hyub Kim, Senay Kafkas, Johanna McEntyre

EMBL-EBI, United Kingdom

Presenter: Jee-Hyub Kim

As the number of research articles in biomedical domain significantly increases, text-mining becomes an attractive approach to help biocurators with finding relevant articles, identifying gene/protein names in the articles, and linking them to biological databases. For this linking, named entity recognition (NER) has been used as a core method. Although NER has been successful, this method often suffers from the ambiguity problem of mined gene/protein names. As an alternative, accession number mining can be used to link articles to databases with high precision. Although a number of publishers provide accession numbers marked up in full-text articles, most accession numbers still remain untagged in free text. In order to mine them, we have developed an accession number tagger. The accession number tagger consists of three steps: 1) Analysing the context of an article, 2) Identifying accession numbers in each sentence and 3) Validating the mined accession numbers using EBI site-wide webservice. Currently, our accession number tagger covers 10 different databases (ENA, UniProt, InterPro, Pfam, PDBe, OMIM, ArrayExpress, Ensembl, refSNP, and refSeq) and data DOIs. With the accession number tagger, we have mined a large number of accession numbers from full text Open Access articles available from Europe PMC. In order to provide mined accession numbers, we have implemented a web service which includes mined accession numbers as well as any other methods useful for biocuration (e.g., retrieving full text articles). Especially, a method (getTextMinedTerms) is designed to provide a list of accession numbers given a PMCID or PMID. Currently, this webservice is available as SOAP and RESTful webservices. In the future, we plan to provide RSS for each database whenever its accession numbers are mentioned in articles. We believe that different access methods of accession numbers will help biocurators with easier uses for their customised annotation tasks.

102

One document, many users: what happens when you re-purpose a document?

David King, David Morse

The Open University, United Kingdom

Presenter: David King

To assess global challenges surrounding issues such as climate change and invasive species requires a baseline of historical data. We are fortunate in biodiversity that such data exists in a rich body of literature. One such source of historical data is the Biologia Centrali-Americana (BCA), which documents the plant and animal life in Central America one hundred years' ago, and which can be compared to contemporary species distributions. This valuable resource has recently been re-keyed and manually marked up by the INOTAXA project (http://www.inotaxa.org/). The 56-volume work is now being curated before wider release. The manual annotation of the BCA is both time consuming in its initial phases and demands expert review to curate the results. This manual approach to mining historic texts is not viable for large-scale works such as the BCA. Attempts to automate the process face the problem of not having suitable corpora against which to develop and then test machine learning tools such as text mining. One project, ViBRANT (http://vbrant.eu/), sought to use the scale of the re-keyed data being produced by INOTAXA to develop a corpus that could be used for this purpose. However, the apparently straightforward task of annotating the BCA has highlighted many issues because different audiences have different requirements of the mark up. This presentation describes the process by which the BCA is being reworked from digitisation through to a curated document corpus. The intended users are biodiversity scientists who can use the corpus for taxonomic and biodiversity research, and computer scientists who can use it to develop new text mining and mark up tools. The presentation covers the different requirements of scientists in the two domains, how this affects the mark up required of the documents, and how to re-purpose the annotations to meet the needs of different and sometimes disparate scientific audiences.

103

Whole Cancer Genome Annotation in the Catalogue Of Somatic Mutations In Cancer (COSMIC)

Chai Yin Kok[1], Simon Forbes[2], John Gamble[2], Rebecca Shepherd[2], Sally Bamford[2], Charlotte Cole[2], S Ward[2], Mingmin Jia[2], Kenric Leung[2], Andrew Menzies[2], Nidhi Bindal[2], David Beare[2], Prasad Gunasekaran[2], Adam P Butler[2], Jon W Teague[2], Peter J Campbell[2], Michael Stratton[2]

[1] Cancer Genome Project, United Kingdom
2 Wellcome Trust Sanger Institute, United Kingdom

Presenter: Chai Yin Kok

COSMIC, the Catalogue Of Somatic Mutations In Cancer (www.sanger.ac.uk/cosmic) is designed to store and display somatic mutation information relating to human cancers, combining detailed information on publications, samples and mutation types. The information is curated both from the primary literature, international consortium data portals and the laboratories at the Cancer Genome Project, Sanger Institute, UK. We have been focusing on annotating cancer data related to high throughput whole genome cancer from publications since 2008. In the recent v63 release, whole-genome annotations were curated across 4, 653 tumour samples, detailing 607, 6 19 mutations in 24, 517 genes, together with 7584 genomic rearrangements. This curation currently comprises 100 genome-wide screens, including 13 cancer types from the two main cancer portals, the International Cancer Genome Consortium (ICGC) and The Cancer Genome Atlas (TCGA). Our curation process gathers genomic mutation details from these sources and generates gene annotations from GRCh37 genomic co-ords through our annotation pipeline, VAGRENT (http://www.sanger.ac.uk/resources/software/vagrent/). This pipeline annotates simple and complex somatic mutations to Human Genome Variation Society (HGVS) nomenclature standards, including quality checking procedures to ensure COSMIC presents the publication data accurately. The data within COSMIC is freely available without restriction, primarily via an analytical website. This system provides users with interactive features to query these annotations by tumour site, histology or gene, and displays the information in graphical and tabular forms, allowing custom exports in various formats.

COSMIC data is also available for full and free download, in spreadsheet and database formats (ftp://ftp.sanger.ac.uk/pub/CGP/cosmic) and an independent Biomart system.

104

Atlas of Cancer Signalling Networks (ACSN): an Institut Curie systems biology platform for molecular cancer research

Inna Kuperstein[1], David Cohen[1], Hien-Anh Nguyen[1], Simon Fourquet[1], Laurence Calzone[2], Stuart Pook[1], Inna Kuperstein[1], Paola Vera-Licona[1], Eric Bonnet[1], Andrei Zinovyev[1], Emmanuel Barillot[1]

[1] Institut Curie, France
[2] U900 INSERM - Mines ParisTech - Institut Curie, France

Presenter: Inna Kuperstein

ACSN (http://acsn.curie.fr) is a platform that contains interconnected cancer-related signalling network maps amenable for computational analysis. Cell signalling mechanisms are depicted on the maps in great detail, demonstrating interactions between cell processes together creating a 'geographic-like' map of molecular interactions in cancer. Signalling maps in ACSN are manually constructed based on the literature describing biochemical mechanisms, using the CellDesigner tool and the Systems Biology Graphical Notation (SBGN). ACSN maps are interconnected into a global cancer signalling network represented in a hierarchical manner. Canonical signalling pathways are indicated and can be visualized as modules. The modular structure of the ACSN maps facilitates navigation through maps. Currently ACSN consists of four maps: cell-cycle regulation by RB-E2F, DNA repair, cell cycle and checkpoints, apoptosis and energy metabolism and cell survival signalling networks. We will extend the ACSN additional maps for epithelial-mesenchymal transition (EMT), telomeres and centrosome maintenance, DNA replication, inflammatory processes and others. The ACSN map navigation, curation and maintenance are enabled by a user friendly Google Maps-based tool named NaviCell (http://navicell.curie.fr). The tool is characterized by the unique combination of three essential features: (1) map navigation based on Google Maps engine, (2) semantic zooming for viewing different levels of details of the map and (3) integrated web-based blog for collecting the community curation feedbacks. NaviCell facilitates curation of molecular interactions maps by the community helping to update and maintain maps in an interactive and user-friendly fashion. ACSN is a platform that helps to address questions like high-throughput data integration and analysis, modeling of synthetic interactions, finding intervention points and optimal drugable targets.

105

Modeling of Sensing-Response Processes in Escherichia coli: Transcriptional Regulation in its Biological Context

Daniela Ledezma-Tejeida[1], Socorro Gama-Castro[1], Kristof Engelen[2], Santiago Sandoval[1], Julio Collado[3]

[1] UNAM, Mexico
[2] Katholieke Universiteit Leuven, Belgium
[3] RegulonDB, Mexico

Presenter: Daniela Ledezma-Tejeida

Bacteria are able to to survive in a changing environment by sensing fluctuations and orchestrating systemic responses that allow it to adapt to new conditions. Transcription factors (TF's) are the link between the environment and the cell genetic machinery, these proteins sense the intracellular concentration of a specific ligand and accordingly alter the expression levels of a defined set of genes, which in turn alter protein levels. The change in protein activity adjusts metabolism to contend with the initial fluctuation in ligand availability. In this circular process four parts can be identified: (1) the environmental signal, internal or external, (2) the processing of the signal, chemical reactions that turn the signal into the ligand recognized by a TF, (3) the genetic switch, changes in TF conformation that leads to changes in gene expression, and (4) the response, the effect of differential protein levels in metabolism that usually relate to the signal. We have termed this process Genetic Sensory Response Unit or GENSOR Unit (GU). We have assembled GUs for 24 TFs in Escherichia coli K12 whose function is related to amino acid biosynthesis and carbon source utilization. This GU set was humanly curated and is publicly available through the Transcriptional Regulatory Network database RegulonDB. An automatic method was developed to extract data from specialized databases, analyze the role of gene products in metabolism and organize data in a human-readable format. With this tool a second set of 177 GUs has been assembled and is currently under human curation. To further demonstrate the relevance of the GU concept we have fitted a boolean model to mathematically describe the interactions among the elements of the publicly available GUs. Currently, we are identifying the steady states to predict mutant phenotypes. Finally, we will use the expression profile database COLOMBOS to validate our predictions under different conditions.

106

Nematode Annotation in UniProtKB/Swiss-Prot

Duncan Legge

EMBL-EBI, United Kingdom

Presenter: Duncan Legge

UniProt (Universal Protein Resource; http://www.uniprot.org) provides a central resource on protein sequences and functional annotation. The UniProt Knowledgebase (UniProtKB) is divided into two parts; the manually annotated UniProtKB/Swiss-Prot section and the automatically annotated UniProtKB/TrEMBL section. Nematode annotation in UniProtKB focuses on the Caenorhabditis elegans worm with priority given to well-studied mammalian homologs or requests from the research community. UniProtKB also contains entries from other Caenorhabditis species notably; briggsae, drosophilae, japonica, remanei and vulgaris. Once a protein entry has been researched and curated, it can contain a variety of information using controlled vocabularies and free text including taxonomy, enzyme- activity, domains and sites, post-translational modifications, tissue- and developmental- specific expression, interactions, splice isoforms, and sequence conflicts. Entries connect to various external data collections such as nucleotide databases, protein structure databases, protein domain and family databases, ontologies, and species- and function-specific data collections. As a result, UniProtKB acts as a central hub connecting biomolecular information, providing a centralized and trusted resource. UniProt provides a mapping service to convert common gene and protein identifiers to UniProtKB accessions/IDs and vice versa. UniProt data is freely available and releases, which happen every four weeks, are provided in Fasta, flat file and XML formats. With ongoing curation of Caenorhabditis elegans proteins, UniProt aims to support the worm research community. Furthermore, we welcome feedback and would be interested to hear which protein families or molecular pathways are of particular interest to your work.

107

SEQwiki: an extensive community created database of tools for high-throughput sequencing.

Jing-Woei Li[1], Keith Robinson[2], Marcel Martin[3], Andreas Sjödin[4], Björn Usadel[5], Matthew Young[5], Eric Olivares[6], Dan Bolser[7]

[1] The Chinese University of Hong Kong, Hong Kong
[2] Warp Drive Biosynthetics, United States of America
[3] Bioinformatics for High-throughput Technologies, Germany
[4] Division of CBRN Security and Defence, Sweden
[5] Max Planck Institute of Molecular Plant Physiology, Germany
[6] SEQanswers.com, United States of America
[7] EMBL-EBI, United Kingdom

Presenter: Dan Bolser

SEQwiki (http://SEQwiki.org) is a wiki database that is actively edited and updated by the members of the SEQanswers community (http://SEQanswers.com). The wiki provides an extensive catalogue of tools and technologies for high-throughput sequencing (HTS), including information about HTS service providers. It has been implemented in MediaWiki with the Semantic MediaWiki and Semantic Forms extensions to collect structured data, providing powerful navigation and reporting features. Within 2 years, the community has created pages for over 500 tools, with approximately 400 literature references and 600 web links. This collaborative effort has made SEQwiki the most comprehensive database of HTS tools anywhere on the web. Here we provide a detailed analysis of the initial and continued growth of this highly successful community annotation project. Like similar projects. We see a model whereby significant increases in content are provided by an increasing number of less active contributors.

108

Manual Annotation of alternative Poly-A sites in the zebrafish genome

David Lloyd, Gavin Laird, Sarah Donaldson, Charles Steward

Wellcome Trust Sanger Institute, United Kingdom

Presenter: David Lloyd

The zebrafish genome is being sequenced and analysed in its entirety at the Wellcome Trust Sanger Institute (WTSI.) Work is on-going to improve the zebrafish genome assembly through the Genome Reference Consortium (www.ncbi.nlm.nih.gov/projects/genome/assembly/grc) using a combination of clone finishing and whole genome data, with a new assembly (Zv10) planned by the end of 2013. To be of most use the zebrafish genome must be annotated with all the biologically functional genomic elements. Automatic annotation provides a quick method for identifying at least one coding variant per locus, however manual annotation is the gold standard for accurate and comprehensive annotation. This is provided by the Human and Vertebrate Analysis and Annotation group (HAVANA) and is available in Vega. (vega.sanger.ac.uk) The combined manual and automatic dataset is available in Ensembl. (www.ensembl.org) Annotation of the zebrafish genome has been restricted by limited data compared to the Human genome which has led to incomplete or fragmented loci. The WTSI has produced RNAseq data (PMID:20081834) to aid manual annotation of the latest reference assembly, Zv9. The resulting gene models give new evidence for exon and intron structure of loci, including non-coding genes, complex gene structures, clusters and rearrangements across a range of different development stages and tissue types. The recent development of 3P-seq data (PMID:22722342) has introduced a large number of alternative polyadenalyation sites, leading to significant improvements in the quality, quantity and completeness of annotated genes in Vega. Our annotation is completed in collaboration with the Zebrafish Information Network (www.zfin.org) building an accurate, dynamic and distinct resource for the zebrafish community, and contributes to the Zebrafish knock-out project (www.sanger.ac.uk/Projects/D_rerio/zmp) Any regions where manual annotation would be of interest can be communicated to zfish-help@sanger.ac.uk.

109

Genomic data and metadata integration for the Saccharomyces Genome Database

Paul Lloyd[1], J. Michael Cherry[1], SGD Project[2]

[1] Stanford University, United States of America
[2] The Saccharomyces Genome Database, United States of America

Presenter: Paul Lloyd

The Saccharomyces Genome Database (SGD) is a comprehensive resource for curated, molecular and genetic information on the genes and proteins of S. cerevisiae. Using a pipeline parallel to the SGD standard curation review process, we are building a diverse library of genome wide datasets generated by current high throughput technologies. In combination with the existing SGD curated data and infrastructure, this provides an easily accessible yeast genomics resource, already featuring over 100 published datasets. In addition to the original published data, the SGD is providing visualization of these genome wide yeast data sets as GBrowse tracks at www.yeastgenome.org. The selected datasets include analyses and processed data from a wide range of landmark yeast publications and are directly comparable to the modENCODE and ENCODE projects. An essential component of these major genomic projects is the associated metadata, which facilitates accessibility and comparison across a huge range of biological factors. Our updated processing pipeline has enabled us to more than double the quantity of metadata collected for each dataset and enhances search functions to facilitate easier access for researchers seeking high quality yeast genomic data. The collected metadata is also structured using the same formats, ontologies and controlled vocabularies as the latest phase of the ENCODE project. The metadata structure is flexible for several downstream applications in the data processing pipeline. Using standard templates, we are able to automatically generate configuration files for genome visualization tools and headers for associated data files. All metadata terms collected are now searchable using a faceted browsing display, results of which link to both GBrowse tracks and downloadable datasets of the associated data.

110

A configurable tool for community literature curation

Antonia Lock[1], Kim Rutherford[2], Midori Harris[2], Stephen Oliver[2], Jürg Bähler[1], Valerie Wood[2]

[1] UCL / PomBase, United Kingdom
[2] University of Cambridge, United Kingdom

Presenter: Antonia Lock

We have developed a web-based annotation tool, Canto, to support community curation on a large scale. Canto is highly configurable, and can be used with minimal or extensive support from professional curators. It is therefore suitable for use by most research communities, including those not supported by a manual curation team, who want to contribute gene-specific experimental information from their organism of interest to public biological databases. Canto supports literature-based curation of a wide, and configurable, set of data types, including Gene Ontology (GO) annotations, phenotypes, interactions, and protein modifications. The tool is fully accessible online, requiring no software download or setup by the end user. Initial feedback from early community users indicates that Canto is easy to use, with an intuitive workflow and integrated help documentation. Canto was originally developed for community curation by the S. pombe database (PomBase) and its research community, who curate the most extensive set of data types. To date, Canto has also been adopted by the K. pastoris (Pichia) community and for GO annotation workshops at University College London in which researchers and post-graduates are invited to curate their own papers of interest. Ongoing Canto development ensures that feedback from users guides efforts to improve existing features or to implement new ones.

111

The impact of focused Gene Ontology annotation efforts on transcriptomic and proteomic data analysis

Ruth Lovering[1], Varsha Khodiyar[1], Rachael Huntley[2], Yasmin Alam-Faruque[2], Emily Dimmer[2], Tony Sawford[2], Maria-Jesus Martin[2], Claire O'Donovan[2], Rolf Apweiler[2], Philippa Talmud[1]

[1] University College London, United Kingdom
[2] EMBL-EBI, United Kingdom

Presenter: Ruth Lovering

Over the last 5 years, the Cardiovascular Gene Ontology (GO) Annotation Initiative, funded by the British Heart Foundation (BHF) and based at UCL, has supplied GO annotation specifically to human proteins involved in cardiovascular processes. This was one of the first annotation efforts to focus on a specific area of biology and working at UCL has enabled the BHF-funded GO curators to establish collaborations with local cardiovascular researchers and build up their own expertise in this area. This has fed back into the further development of the Gene Ontology itself, with the UCL curation team being responsible for the creation of ~1600 GO terms. Some of these terms have been created as the result of concerted ontology development efforts, such as terms to describe heart development; others have been created on an ad-hoc basis as needed. The interpretation of transcriptomic and proteomic analyses is often limited by the quality and quantity of the annotations available. Our analyses of a hypertension transcriptomic dataset demonstrated that the ability to identify discriminatory groupings can be vastly improved by combining the creation of more specific GO terms with the use of these terms to provide more descriptive protein functional annotations. Removing the BHF-funded annotations from this analysis decreased the significance of the majority of enriched GO terms and several disease-relevant GO terms were no longer identified, such as 'cytokine-mediated signaling pathway'. Our analysis demonstrates the need for freely available functional analysis tools to regularly update the GO datasets incorporated into these tools.

Supported by British Heart Foundation grant SP/07/007/23671.

112

InterMine: Interoperation and Embeddable Analysis Tools

Rachel Lyne, Jelena Aleksic, Daniela Butano, Sergio Contrino, Hu Fengyuan, Alex Kalderimis, Mike Lyne, Radek Stepan, Julie Sullivan, Gos Micklem

University of Cambridge, United Kingdom

Presenter: Rachel Lyne

InterMine is an open source data warehouse built specifically for the integration and analysis of complex biological data. InterMine enables the creation of biological databases accessed by sophisticated web query tools and web services. Five of the major Model Organism Databases (MODs), yeast, rat, mouse, fish and nematode are adopting the InterMine data integration platform to provide flexible searching and data mining interfaces to their user communities. The InterMOD project aims to facilitate interoperation between these MOD InterMines, thus making it easier for users to navigate between MOD databases and enabling cross-data comparisons between organisms. The InterMine framework aims to include a user-friendly web interface as well as a powerful, scriptable web service API to allow programmatic access to data. The web interface includes a useful identifier resolution system, sophisticated query options and interactive results tables that enable powerful exploration of data, including data summaries, filtering, browsing and export to external analysis in Galaxy. A set of graphical analysis tools provide a rich environment for data exploration including statistical enrichment of sets of genes or other biological entities. The embeddable nature of the results tables and analysis tools means their functionality can be placed on any webpage with minimal development overhead, for example allowing data from different organisms to be viewed on a single webpage. The use of InterMine as a cross-species comparison platform and its library of reusable web parts will be presented.

113

Leveraging ontologies for experimental discovery at the ENCODE Portal

Venkat Malladi[1], Eurie Hong[1], Galt Barber[2], Gail Binkley[1], Esther Chan[1], Drew Erickson[1], Benjamin Hitz[3], Donna Karolchik[2], Katrina Learned[2], Brian Lee[2], Jeffrey Long[1], Kate Rosenbloom[2], Greg Roe[1], Laurence Rowe[1], Cricket Sloan[1], J Strattan[1], William Kent[2], J. Michael Cherry[1]

[1] Stanford University, United States of America
[2] University of California Santa Cruz (UCSC), United States of America
[3] SGD, United States of America

Presenter: Venkat Malladi

The Encyclopedia of DNA Elements (ENCODE) Project is a collaborative project to create a comprehensive catalog of functional elements in the human and mouse genomes. To date, the ENCODE project has generated 3000+ experiments, using over 30 different experimental techniques and 300+ cell lines and tissue types, in order to investigate the binding sites of over 200 DNA-binding proteins, refine the annotation of protein-coding genes and non-coding RNAs, and examine the chromatin structure. As the ENCODE project enters its next phase of production, the data volume is expected to increase, covering more DNA-binding proteins, chromatin structure, and transcription in more cell lines and tissue types. All data generated by the ENCODE project are submitted to the Data Coordination Center (DCC) for validation, tracking, storage, distribution, and visualization to community resources and the scientific community. The number of data sets and the complexity of the methods used makes identification of experiments that match the interests of a researcher challenging. To accommodate the complex and diverse needs of researchers a new facility is being created. The ability to identify and download the appropriate data is being enhanced with the expansion of metadata that will be used to describe an experiment, the use of bio-ontologies to annotate these metadata, and the development of ontology-driven search tools that will facilitate identification of these data. The DCC plans to utilize the Cell Type Ontology (http://cellontology.org/) and UBERON (http://uberon.org/) to facilitate identification of high-throughput data generated from common anatomical structures, cellular morphologies, or developmental stages using a single search entry point.

Data from the ENCODE project can be accessed via the ENCODE portal (http://www.encodeproject.org) and the UCSC Genome Browser (http://genome.ucsc.edu/cgi-bin/hgGateway).

114

The intricacies of mouse genome annotation

Deepa Manthravadi, If Barnes, Marie Marthe-Suner, Jonathan Mudge, Charles Steward, Mark Thomas, Laurens Wilming, Jennifer Harrow

Wellcome Trust Sanger Institute, United Kingdom

Presenter: Deepa Manthravadi

The HAVANA group provides the scientific community with high quality manual annotation of mouse, a model organism in many fields such as immunology, genetics, genomics and molecular biology. In order to do this, we have annotated several challenging regions in the mouse genome, for example the α takusan and the α and β defensin gene clusters. These highly repetitive genes exist in multiple duplicated copies, are subject to haplotypic variation and are difficult to classify. In collaboration with the Mouse Genome Informatics group we provide consistent nomenclature for these and other gene clusters. Genes from the vomeronasal receptor, olfactory receptor and major urinary protein (MUP) families are also located in large clusters, greatly expanded in mouse compared to human. MUPs also vary between two different mouse strains we have manually annotated. We observed divergence of gene clusters between mouse and human prolactin and serine protease inhibitor genes, with greater expansion of these clusters in mouse than human, and the reverse for growth hormone genes. The Major Histocompatibility Complex region has also benefited from comparative manual annotation, in C57BL/6 reference and several other strains. Via the Consensus Coding Sequence project we contribute to the generation of a consensus coding gene set and give feedback to the Genome Reference Consortium to ensure the quality of the mouse genome assembly. The HAVANA group is a member of the International Knockout Mouse Consortium whose aim is to design mouse knockout cell lines for every coding gene. This requires first creating a comprehensive, accurate genome annotation and then creating vector designs for knockouts. In this presentation we will showcase the advantages of manual genome annotation in mouse for complex gene regions.

All our annotation is available through the VEGA genome browser (vega.sanger.ac.uk).

115

Improving the FlyBase bibliography

Steven Marygold, Paul Leyland, Ruth Seal, Joshua Goodman, Jim Thurmond, Victor Strelets, Robert Wilson, FlyBase Consortium

FlyBase, United Kingdom

Presenter: Steven Marygold

An accurate, comprehensive, non-redundant and up-to-date bibliography is a crucial component of any Model Organism Database (MOD). Principally, the bibliography provides a set of references that are specific to the field served by the MOD. Moreover, it serves as a backbone to which all curated biological data can be attributed. Here, we describe the organization and main features of the bibliography in FlyBase (flybase.org), the MOD for Drosophila melanogaster. We present an overview of the current content of the bibliography, the pipeline for identifying and adding new references, the presentation of data within Reference Reports and effective methods for searching and retrieving bibliographic data. We highlight recent improvements in these areas and describe the advantages of using the FlyBase bibliography over alternative literature resources.

116

HBV interactive replication cycle in ViralZone

Patrick Masson[1], Philippe Le Mercier[1], Chantal Hulo[1], De Castro Edouard[1], Hans Bitter[2], Lore Gruenbaum[2], Laurent Essioux[3], Lydie Bougueleret[1], Ioannis Xenarios[1]

[1] SIB Swiss Institute of Bioinformatics, Swiss-Prot Group, Switzerland
[2] Translational Research Sciences, Hoffmann-La Roche Drug Discovery & Early Development, United States of America
[3] Hoffmann La Roche, Bioinformatics and Exploratory Statistics Department, pRED, Switzerland

Presenter: Patrick Masson

Hepatitis B is a viral infection that attacks the liver and can cause both acute and chronic disease. Two billion people worldwide have been infected with the virus and about 600 000 people die every year due to the consequences of hepatitis B virus (HBV) infection. The virus genome is surprisingly small, only 3.2kb long, and encodes seven overlapping proteins. Every base pair in the genome is involved in coding at least one viral protein! Despite its apparent simplicity, HBV life cycle is complex and still subject of intensive research. In partnership with Hoffmann-La Roche we created an interactive HBV life cycle resource that can be accessed through the ViralZone portal (http://viralzone.expasy.org/all_by_protein/1280.html) . This resource gathers knowledge about HBV cellular infection in a single place. The entry point to the HBV cycle is an illustration depicting the virus replication cycle in a hepatocyte host. The cycle has been divided into 27 steps and molecular events, each linking to HBV specific description pages. Most of these steps are described by terms from the viral ontology developed for the UniProtKB virus annotation program. Each description page contains a synthesis of the current knowledge and is associated with all major publications and annotation comments.

117
Crop Ontology, a vocabulary for crop-related concepts curated by the community

Luca Matteis, Elizabeth Arnaud

Bioversity International, France

Presenter: Luca Matteis

The Crop Ontology curation website allows crop-related data to be part of the Linked Data repository, by building and sharing a common vocabulary of crop-related concepts that is maintained and curated by the community. Linked Data, http://linkeddata.org/, is the largest initiative on the web that tries to aggregate data from a variety of different resources. It does this through standard semantic tools and technologies such as RDF (Resource Description Framework), which allows datasets to agree on common schemas. We show how data can be annotated using the Crop Ontology, and shared on the web so that it can be easily queried using standard web semantic tools.

118

The HUPO Proteomics Standards Initiative-Mass Spectrometry Controlled Vocabulary

Gerhard Mayer[1], Luisa Montecchi-Palazzi[2], David Ovelleiro[2], Andrew Jones[3], Pierre-Alain Binz[4], Eric Deutsch[5], Matthew Chambers[6], Marius Kallhardt[7], Fredrik Levander[8], James Shofstahl[9], Sandra Orchard[2], Juan Antonio Vizcaino[2], Henning Hermjakob[2], Christian Stephan[10], Helmut Meyer[11], Martin Eisenacher[12]

[1] Ruhr-Universität Bochum, Germany
[2] EMBL-EBI, United Kingdom
[3] Institute of Integrative Biology, University of Liverpool, United Kingdom
[4] SIB / GeneBio, Switzerland
[5] Institute for Systems Biology, United States of America
[6] Department of Biomedical Informatics, Vanderbilt University Medical Center, United States of America
[7] Bruker Daltonik GmbH, Germany
[8] BILS, Department of Immunotechnology, Lund University, Sweden
[9] Thermo Fisher Scientific Inc., United States of America
[10] Kairos GmbH, Germany
[11] Medizinisches Proteom Center (MPC), Germany
[12] Medizinisches Proteom Center (MPC), Ruhr-Universität Bochum, Germany

Presenter: Gerhard Mayer

Controlled vocabularies (CVs) and ontologies are used in structured data formats and databases to avoid inconsistencies in annotation, to have a unique (and preferably short) accession number and to give researchers and computer algorithms the possibility for more expressive semantic annotation of data. The Human Proteome Organization (HUPO) - Proteomics Standards Initiative (PSI) makes extensive use of ontologies / CVs in their data formats. The PSI-Mass Spectrometry (MS) controlled vocabulary contains all the terms used in the PSI mass spectrometry-related data standards. The CV contains a logical hierarchical structure to ensure ease of maintenance and the development of software that makes use of complex semantics. The CV contains terms required for a complete description of a mass spectrometry analysis pipeline used in proteomics, including sample labeling, digestion enzymes, instrumentation parts and parameters, software used for identification and quantification of peptides/proteins and the parameters and scores used to determine their significance. Due to the range of topics covered by the CV, collaborative development across several PSI working groups, including proteomics research groups, instrument manufacturers and software vendors, was necessary. In this article, we describe the overall structure of the CV, the process by which it has been developed and is maintained, and the dependencies on other ontologies.

119

Using text-mining to streamline the Literature curation process in FlyBase.

Peter McQuilton[1], Yuling Li[2], Juan-Miguel Cejuela[3], Hans-Michael Muller[2], Paul Sternberg[2], Burkhard Rost[3], Nicholas H. Brown[1]

[1] University of Cambridge, United Kingdom
[2] California Institute of Technology, United States of America
[3] Technische Universität München, Germany

Presenter: Peter McQuilton

FlyBase (www.flybase.org) is the premier database for Drosophila genetic and genomic information. Over the last 20 years, FlyBase has had to adapt and change to keep abreast of advances in biology and database design and we are continually looking for ways to improve curation efficiency and efficacy. Genetic literature curation focuses on the extraction of genetic entities (e.g. genes, mutant alleles, transgenic constructs) and their associated phenotypes and Gene Ontology terms from the published literature. Over 2000 Drosophila research articles are now published every year and these articles are becoming ever more data-rich. There is a growing need for text mining to shoulder some of the burden of paper triage and data extraction. In this poster, we describe our curation workflow, along with some of the problems and bottlenecks therein, and highlight the ways in which we are incorporating text-mining into the workflow to reduce the curation burden. We will describe our use of Support Vector Machine methods (in collaboration with the WormBase Textpresso group) to aid the triage of new papers based on the presence of particular data-types, such as the presence of a new allele or new transgene, and our use of machine-learning (in collaboration with the Rost lab) to identify gene symbols in the results and materials and methods sections of a paper to aid curation.

120

CSFGBro, a Browser Companion for Biocurators of Fungal Genomics Literature

Vahe Chahinian, Marie-Jean Meurs, Erin McDonnell, Ingo Morgenstern, Greg Butler, Adrian Tsang

Centre for Structural and Functional Genomics, Concordia University, Canada

Presenter: Marie-Jean Meurs

Discovery and development of effective fungal enzyme cocktails are cornerstones of the biorefinery industry because these cocktails can convert lignocellulose into fermentable sugars for biofuel production. The manual curation of fungal genes encoding lignocellulose-active enzymes is an essential step for supporting further research and experiments, as it allows researchers to easily access reliable knowledge. We present CSFGBro, an augmented browsing tool supporting the manual curation of literature related to genomics-based lignocellulose research. In Web pages, CSFGBro highlights information automatically retrieved by the mycoMINE text mining system. Developed in close collaboration with researchers working on the Genozymes [http://www.fungalgenomics.ca] project, mycoMINE annotates documents with tags such as enzyme, fungus, activity or pH/Temperature conditions. Tag categories of interest have been selected according to those reported in the mycoCLAP database [http://mycoclap.fungalgenomics.ca]. CSFGBro provides biocurators with an overview of the document content. A sidebar displays mentions retrieved in the document. Clicking on a highlighted mention opens a popup, which shows the features of the mention, any available standard identifiers, and hyperlinks to external sources of knowledge. For instance, an enzyme popup displays the enzyme EC Number, its recommended and systematic names reported on BRENDA, its SwissProt identifiers, and hyperlinks to related pages. The CSFGBro backend is composed of a RESTful interface programmed with JavaScript [http://nodejs.org/]. The interface highlights mycoMINE results, creates a sidebar menu listing them, and adds the on-click functionality. The CSFGBro front end is a user script that sends requests to the interface to retrieve the data and then displays it. Two curators evaluated a prototype with limited features on the triage of 114 PubMed abstracts. Using the tool, the time needed for triage was reduced by 21%.

121

Supporting Triage of PubMed Abstracts for mycoCLAP

Marie-Jean Meurs, Erin McDonnell, Ingo Morgenstern, Greg Butler, Justin Powlowski, Adrian Tsang

Centre for Structural and Functional Genomics, Concordia University, Canada

Presenter: Marie-Jean Meurs

Fungi secrete a variety of enzymes that work efficiently to degrade lignocellulosic biomass. Since the breakdown of lignocellulose is vital for a number of industrial processes that stand to be improved, interest in these enzymes is high. Fungal lignocellulose-degrading enzymes are numerous and display a wide range of characteristics and properties. As such, the curation of those that have been characterized is essential for supporting further research into their potential applications. To date, several types of fungal lignocellulose-degrading enzymes have been manually curated and deposited into mycoCALP, a searchable database of Characterized Lignocellulose-Active Proteins of Fungal Origin [http://mycoclap.fungalgenomics.ca]. The curation process for mycoCLAP involves a number of steps. The most time-consuming one involves finding appropriate papers to curate. Literature searches often recover many candidate articles which then need to be manually screened for relevance by the curator (triage task). A reliable automated screening and filtering approach of the candidate articles would save precious time. The work we present supports the automatic triage of PubMed abstract candidates. The developed processing resource takes PubMed abstracts as input, then classifies them according to their potential for full paper curation for mycoCLAP. The inference engine making the classification decision is based on first order logic rules combining constraints based on the document topic with constraints based on presence of entities or concepts in smaller units of text. A first evaluation was performed on 104 PubMed abstracts published from 11.01.2012 to 01.30.2013 and retrieved by keyword search for fungal oxidoreductase; lignin, versatile and manganese peroxidase; pyranose oxidase; glyoxal oxidase. The system selection was checked against manual triage, validating our approach with these results: precision=0.68, recall=0.79, true negative rate=0.83, and accuracy=0.88.

122

Strength of IMGT® standards in NGS repertoire analysis of IG and TR with IMGT/HighV-QUEST

Joumana Michaloud, Géraldine Folch, Véronique Giudicelli, Patrice Duroux, Eltaf Alamyar, Marie-Paule Lefranc

IMGT, the international ImMunoGeneTics information system®, Laboratoire d'ImmunoGénétique Moléculaire LIGM, Institut de Génétique Humaine IGH, UPR CNRS, France

Presenter: Joumana Michaloud

The analysis of expressed repertoires of immunoglobulins (IG) and T cell receptors (TR) represents a huge challenge for the study of the adaptive immune response in normal and disease-related situations. To answer that need IMGT®, the international ImMunoGeneTics information system® has developed IMGT/HighV-QUEST [1,2] for the analysis of large repertoires of IG and TR sequences from NGS, which analyses up to 150,000 sequences per run, and provides statistical analysis for up to 450,000 sequences. IMGT/HighV-QUEST identifies the V, D, J genes and alleles by alignment with the germline IG and TR gene and allele sequences of the IMGT reference directory, which is constructed with data resulting from IMGT expert annotation. IMGT/HighV-QUEST integrates IMGT/JunctionAnalysis for a detailed analysis of the V-J and V-D-J junctions, and IMGT/Automat for a full V-J and V-D-J annotation. This analysis is based on IMGT-ONTOLOGY [3], which includes seven axioms: IDENTIFICATION, DESCRIPTION, CLASSIFICATION, NUMEROTATION, LOCALIZATION, ORIENTATION and OBTENTION. These axioms have led to the generation of concepts and, based on these concepts, to the IMGT Scientific chart rules and to IMGT standards: thus, for examples, the concepts of identification have led to the IMGT keywords, the concepts of description to the IMGT labels, the concepts of classification to the IMGT nomenclature and the concepts of numerotation to the IMGT unique numbering and to the IMGT Colliers de Perles used for antibody engineering and humanization. IMGT standards, the basis of IMGT biocuration, allow IMGT/HighV-QUEST to analyse NGS sequences of the expressed repertoires of antigen receptors with the same degree of accuracy and detailed annotation (539 columns) as IMGT/V-QUEST.

Since October 2010, 270 millions of sequences from 461 users have been analysed. [1] Alamyar E et al. Mol Biol 882:569-604, 2012. [2] Alamyar E et al. Immunome Res8(1):26, 2012. [3] Giudicelli V and Lefranc M-P. Front Genet 3:79, 2012.

123

Rhea, a comprehensive resource of expert-curated biochemical reactions

Anne Morgat[1], Kristian Axelsen[1], Rafael Alcantara[2], Eugeni Belda[3], Elisabeth Coudert[1], Maria-José Lopez Sanchez[3], Marcus Ennis[2], Steve Turner[2], Janna Hastings[2], Lydie Bougueleret[1], Ioannis Xenarios[1], Alan Bridge[1], Christoph Steinbeck[2]

[1] SIB Swiss Institute of Bioinformatics, Switzerland
[2] EMBL-EBI, United Kingdom
[3] Genoscope—LABGeM, CEA, France

Presenter: Anne Morgat

Rhea (http://www.ebi.ac.uk/rhea) provides a non-redundant set of chemical transformations for use in a broad spectrum of applications, including metabolic network reconstruction and pathway inference. Rhea includes enzyme-catalyzed reactions (covering the IUBMB Enzyme Nomenclature list), transport reactions and spontaneously occurring reactions. Rhea reactions are described using chemical species from the ChEBI resource (http://www.ebi.ac.uk/chebi) and are stoichiometrically balanced for mass and charge. They are extensively manually curated with links to source literature and other public resources on metabolism including enzyme and pathway databases. This cross-referencing facilitates the mapping and reconciliation of common reactions and compounds between distinct resources, which is a common first step in the reconstruction of genome scale metabolic networks and models.

124

PLAIN – A Computational API for Plant Genomic Data

Robert Muller, David Huang, Cynthia Lee, Donghui Li, Eva Huala

Carnegie Institution for Science, United States of America

Presenter: Donghui Li

The Arabidopsis Information Resource (TAIR) provides foundational data and services to the plant biology research community, including the Arabidopsis thaliana reference genome and its functional and structural annotation. The current web site makes the data available through search pages and graphical genome browser displays. The PLAIN project adds several open-source computational application programming interfaces (APIs) to the TAIR tool set to enable computational biologists to access TAIR data easily and efficiently. Based on a completely new object-oriented data warehouse design, PLAIN provides SOAP and REST web services for common data sets, available through the TAIR web site, and the ability to develop customized web services using Model Driven Architecture (MDA) and the Unified Modeling Language (UML). PLAIN bases its queries on Plant/SQL, a version of the ANSI SQL 200x SELECT statement standard, extended with special-purpose expressions to handle queries that are difficult or impossible to express in basic SQL: ontology tree queries, polymorphism-between-variant queries, and polymorphism range queries. PLAIN implements a reference query tool, the TAIR Query Builder, which integrates these technologies in a simple web-based interface. All PLAIN software will be available through github as open source software.

125

Functional profiling of gene and disease features using MeSH controlled vocabulary

Takeru Nakazato

Database Center for Life Science, Japan

Presenter: Takeru Nakazato

The major aims of omics analysis are to understand relationships among diseases and relevant genes. Though we utilize Gene Ontology (GO) to annotate gene features from molecular level, we cannot obtain enough information of associated diseases and drugs by using GO. Additionally, to refer disease information, Online Mendelian Inheritance in Man (OMIM) is a useful resource, but has not been fully exploited for omics analysis because its bibliographic data structure is not suitable for computer automation. Therefore, we generated feature profiles of genes and diseases from associated disease, drugs, and anatomy fields with the Medical Subject Headings (MeSH) controlled vocabulary. MeSH is not assigned to Entrez Gene and OMIM entries directly because it is originally prepared to index MEDLINE articles. So, we generated Entrez Gene/OMIM-MeSH associations by retrieving articles referring to such genes and disease, and extracting assigned MeSH terms from these articles. By comparing feature profiles of types 1 and 2 diabetes, we clearly illustrated their differences: type 1 diabetes is an autoimmune disease (P-value = 4.55 × 10−5) and type 2 diabetes is related to obesity (P-value = 1.18 × 10−15). We developed a web-based application called Gendoo (gene, disease features ontology-based overview system) to visualize these profiles. Gendoo is freely available at http://gendoo.dbcls.jp/. Gendoo can be applied to biological interpretation of gene lists as the results of omics analysis including microarray and Next-generation sequencing (NGS). Also, to find public NGS data relevant to diseases, we attempted to generate hyperlinks between Gendoo diseases and related SRA entries. [Reference] Gendoo: Functional profiling of gene and disease features using MeSH vocabulary, Nakazato, T., Bono, H., Matsuda, H., and Takagi, T., Nucleic Acids Research, 37(Suppl. 2) (Web Server Issue):W166-W169 2009.

126

Mapping the UniProt human reference proteome to the reference genome and variation data

Andrew Nightingale, Jie Luo, Maria-Jesus Martin, UniProt Consortium

EMBL-EBI, United Kingdom

Presenter: Andrew Nightingale

UniProt has annotated the complete Homo sapiens proteome and approximately 20,000 protein coding genes are represented by a canonical protein sequence in UniProtKB/Swiss-Prot. Most of these protein sequences are now mapped to the reference genome assembly produced by the international Genome Reference Consortium (GRC). This mapping was possible through a process of aligning all human protein sequences in the UniProt Knowledgebase (UniProtKB) to the protein translations in Ensembl, based on 100% amino acid identity over the entire sequence. Where protein sequences were not found in UniProtKB, new records were added to the UniProtKB reference proteome data set. As described in a separate poster, UniProt manually annotate variant sequences with known functional consequences from the literature. Studies on sequence variation are increasing; projects like 1000 Genomes and the Cancer Genome Project are generating a vast amount of variant information that is, or will be, stored by Ensembl variation in their databases. This exponential growth of variation information is making it infeasible to expect non-synonymous variants to be only manually assessed and added to UniProtKB. With human reference protein sequences mapped to the reference genome assembly, UniProt has developed a pipeline to import high-quality 1000 Genomes and COSMIC non-synonymous single amino acid variants from Ensembl variation. 389,935 single amino acid variants have been identified for import in the UniProt human reference proteome. This poster describes the mapping procedure and highlights the invaluable information variation data provides to clinical medicine and biological science. UniProt intends to extend this Ensembl variant import pipeline for other species with a complete proteome.

127

Sorting Big Data: Challenges of Capturing and Curating Information in the Mouse Genome Database (MGD)

Hiroaki Onda, MGD Curation Staff, Judith Blake, Carol Bult, Janan Eppig

The Jackson Laboratory, United States of America

Presenter: Hiroaki Onda

The laboratory mouse is widely used as a model for studying normal human biology and disease. The wealth of genetic and genomic resources for mouse, including an extensive genetic map, complete genome sequence, and myriad tools for molecularly manipulating its genome, has caused growth in depth and breadth of biological information about mouse and mouse models for human disease to expand at a dizzying pace. The Mouse Genome Database (MGD; http://www.informatics.jax.org) provides integrated access to these data. Data are obtained from high-throughput data providers, direct data submission from research laboratories, and curated published literature. Integration and normalization of these data sources are an ongoing resource and prioritization challenge. We incorporate and analyze gene models from NCBI, Ensembl, and VEGA, and MGD's curated gene set to maintain a unified gene catalog. We are the authoritative source for mouse Gene Ontology (GO) annotations, combining manually curated functional annotations with those inferred from mouse, human and rat homologs and sequence resources. Phenotypic data derive from experimental studies on a lab-specific basis or from high-throughput phenotyping centers and are curated to unifying Mammalian Phenotype Ontology terms. Curators regularly survey a set of 150 journals for relevant new publications containing MGD-pertinent data in a process termed "literature triage", which has produced a literature corpus of over 185,000 references to date. (see poster from Kevin Cohen and Gully Burns on using NLP to assist with "literature triage"). We will illustrate the versatility of MGD as a research tool, and showcase data integration as a powerful enabler for mining genetic, genomic, phenotypic, and disease model data. We also will discuss challenges and prospects for continued high-quality integration and curation of high-throughput data in MGD. MGD is supported by NIH NHGRI grant HG000330.

128

When do they die? Curating and querying how lethality varies with stage.

David Osumi-Sutherland, Steven Marygold, Laura Ponting, Peter McQuilton, Raymund Stefancsik, Gillian Millburn, Nick Brown

University of Cambridge, United Kingdom

Presenter: David Osumi-Sutherland

Geneticists choosing genotypes for their experiments need to know whether animals will survive to a stage suitable for their chosen experiments. Knowing the various stages at which significant number of mutant animals die can also help researchers to home in on stages to characterise for phenotypic defects. FlyBase records information about the stages of death due to specific genotypes by combining the ontology terms 'lethal' and 'semi-lethal' with terms for stages. However, the terms lethal and semi-lethal have until recently lacked definitions. And, while stage terms are well defined, the ways in which they modify the meaning of 'lethal' and 'semi-lethal' has never been codified. The current formal annotation set in FlyBase is therefore not reliably useful for queries about the stages at which death occurs. We have devised a set of ontology terms, with formal semantics in OWL, for recording and reasoning about the stages at which death occurs. These are defined using terms from the Drosophila stage ontology, along with a formal treatment of penetrance and a set of relations and axioms for reasoning about relative timing. The resulting class hierarchy can be used to query for genotypes that cause all animals to die before some specified stage, or significant numbers of animals to die before or during some specified stage. We will present our formalisation, along with limitations of the OWL modelling approach and examples of how we will use this system in FlyBase curation. We will also outline how our approach may be applicable to other organisms and phenotypes.

129

The bioinformatics.ca Links Directory 2013

David Yim, Winston Leung, Michelle Brazas, Sophia Wen, Joseph Yamada, Francis Ouellette

OICR, Canada

Presenter: Francis Ouellette

The Bioinformatics Links Directory is a compendium of useful bioinformatics links organized in an intuitive hierarchy. Many of the links from the links directory have come from the yearly NAR web Server and Database issues, but it was also supplemented from people in our laboratory and the bioinformatics community at large. This directory was first initiated in 1999, and has been maintained by the people in the Ouellette laboratory at UBC (Vancouver, BC, Canada) and now in Toronto at the Ontario Institute for Cancer Research. In addition to useful bioinformatics links, we also have databases and tools in the Links Directory. More than 90% of all links have one or more PubMed reference, providing more information about the resources. With the migration to Drupal 6 a number of features have been introduced and will facilitate other developments. These include: · Links Directory makes use of web 2.0 and semantic searches · Links Directory includes links, tools, and databases · Ability to facilitate and monitor community input and upkeep · Introduction and management of content and metadata Tags generated from user input and from PubMed MeSH terms for better curation and recall of content. · Development of user groups to encourage collaborate content curation and maintenance within a group or an institution · Web Service/API to get submissions from journal publishers for content diversity We believe that these additions and continued updates will provide and allow a better user experience resulting in finding what you want faster and in the context of other tools and help documentation, all of this in an environment that will be supported by the community and in constant relationship with the key publishers for log-lasting linkage to the key publications. We want to involve the biocuration community in making this a better resource, so please come and visit: http://bioinformatics.ca/links_directory/

130

The BioGRID Interaction Database: New Post-Translational Modification (PTM) Display and Full Coverage of Yeast Interactions

Rose Oughtred[1], Bobby-Joe Breitkreutz[2], Lorrie Boucher[2], Christie Chang[1], Andrew Chatr-Aryamontri[3], Daici Chen[3], Sven Heinicke[1], Jodi Hirschman[1], Nadine Kolas[2], Michael Livstone[1], Julie Nixon[4], Lara O'Donnell[2], Lindsay Ramage[4], Teresa Reguly[2], Jennifer Rust[1], Chris Stark[2], Chandra Theesfeld[1], Andrew Winter[4], Ivan Sadowski[5], Kara Dolinski[1], Mike Tyers[3]

[1] Lewis-Sigler Institute for Integrative Genomics, Princeton University, United States of America
[2] Systems Biology, Samuel Lunenfeld Research Institute, University of Toronto, Canada
[3] Institute for Research in Immunology and Cancer, Université de Montréal, Canada
[4] Wellcome Trust Centre for Cell Biology and Centre for Systems Biology, University of
[5] Department of Biochemistry and Molecular Biology, Molecular Epigenetics, Life Sciences Institute, University of British Columbia, Canada

Presenter: Rose Oughtred

Genetic and physical interaction networks are critical for understanding complex biological systems, and for providing key insights into normal and diseased states. To further this understanding, the Biological General Repository for Interaction Datasets (BioGRID) (http://www.thebiogrid.org) curates and freely disseminates collections of protein and genetic interactions from major model organisms, as well as human. BioGRID currently houses over 620,000 interactions curated from high-throughput datasets and individual studies found in the primary literature. Complete coverage of the entire literature for both the budding yeast Saccharomyces cerevisiae and fission yeast Schizosaccharomyces pombe continues to be maintained, resulting in the curation of over 372,000 interactions culled from more than 13,000 publications. Efforts are underway to curate interactions for human and various model organisms with a focus on disease-related networks such as the Ubiquitin-Proteasome System (UPS), which is implicated in neurodegenerative diseases. New features added to BioGRID include the incorporation of phosphorylation and ubiquitination sites in an integrated post-translational modification (PTM) display. The new PTM viewer contains over 47,000 ubiquitination sites documented by BioGRID on nearly 11,000 human and yeast proteins. More than 20,000 phosphorylation sites on over 3,100 yeast proteins were also manually curated from the S. cerevisiae literature by PhosphoGRID (http://www.phosphogrid.org). BioGRID's interaction data formats continue to be compatible with either the Cytoscape or Osprey network visualization systems. The entire interaction data set may be freely downloaded and source code for BioGRID is also available without any restrictions.

This work is supported by the following grants: NIH RO1RR024031 (PI Mike Tyers), NIH R24RR032659 (PI Kara Dolinski), BBSRC BB/F010486/1 and CIHR CTP-82940 (PI Mike Tyers).

131

The UniRule system in UniProtKB – leveraging manual Swiss-Prot annotation to TrEMBL

Cecilia Arighi[1], Ivo Pedruzzi[2], Klemens Pichler[3], Consortium UniProt[4]

[1] Protein Information Resource, United Kingdom
[2] Swiss Institute of Bioinformatics, United Kingdom
[3] EMBL-EBI, United Kingdom
[4] EMBL-EBI/PIR/SIB, United Kingdom

Presenters: Cecilia Arighi, Ivo Pedruzzi, Klemens Pichler

The UniProt Knowledgebase is a central hub for the collection of functional information on proteins. It consists of two sections, UniProtKB/Swiss-Prot, containing manually-reviewed records with information extracted from the literature and curator-evaluated computational analyses, and UniProtKB/TrEMBL, a section with unreviewed computationally analyzed records. At the time of writing, reviewed entries constitute only 2% of UniProtKB. We have therefore developed the UniRule system that leverages manual curation for the automatic annotation of unreviewed UniProtKB entries. UniRule integrates rules from historically distinct automatic annotation systems (HAMAP, PIRNR/PIRSR, Rulebase) in a single pipeline. UniRule consists of manually created annotation rules that specify functional annotations and the conditions which must be satisfied for them to be applied such as taxonomic scope, family membership as defined by InterPro, and the presence of specific sequence features. Rule generation is part of an integrated workflow that starts with the manual curation of UniProtKB/Swiss-Prot records. The UniRule application is renewed with each UniProt release, and predictions are continuously evaluated against the content of matching reviewed entries, guaranteeing that the predictions remain in synch with the expert curated knowledge of UniProtKB/Swiss-Prot. Currently, UniRule comprises about 3100 rules and provides annotations to 27% of UniProtKB/TrEMBL, greatly extending the impact of manual curation of the UniProt Knowledgebase.

132

The EMMA database: A curation perspective.

Raffaele Matteoni[1], Marzia Massimi[1], Karen Pickford[2], Ann-Marie Mallon[2], Terrence Meehan[3], Gautier Koscielny[3], Michael Hagn[4]

[1] Consiglio Nazionale delle Ricerche, Italy
[2] MRC – Mammalian Genetics Unit, United Kingdom
[3] EMBL-EBI, United Kingdom
[4] Helmholtz Zentrum Munchen, Germany

Presenter: Karen Pickford

The laboratory mouse is an extremely useful model for studying human disease and thousands of mutants have been identified or produced, most recently through gene-specific mutagenesis approaches. The International Knockout Mouse Consortium (IKMC) is now in progress of creating mutants for all protein coding genes, using high throughput strategies. Generating a knock-out line incurs huge financial and time costs, so efficient capture of both the data describing each mutant, alongside archiving of the line for distribution to future researchers is critical. The European Mouse Mutant Archive (EMMA) is a leading international network infrastructure for archiving and worldwide provision of mouse mutant strains, operating as the European component in collaboration with the other members of the Federation of International Mouse Resources (FIMRe). EMMA is also one of four repositories involved in the IKMC, and therefore the current figure of 4300 archived lines is due to increase rapidly. The EMMA database gathers and curates extensive data on each mouse strain and presents it through a user-friendly website. The primary goal of the data curation work is to define and provide gene/allele and strain names/symbols, according to the rules defined by the 'International Standards of Genetic Nomenclature for Mice'. This enables cross-reference with MGI/MGD and IMSR databases and IKMC and related project databases (EUCOMM, EUCOMMTools, IMPC, etc.). Appropriate, up-to-date curation of genetic data and strain nomenclature is also essential for efficient access and use of EMMA-DB resources by requesting scientists. A BioMart interface allows advanced searching including integrated querying with other resources e.g. Ensembl, EuroPhenome, OMIM. Other resources are able to display EMMA data by accessing our Distributed Annotation System server. EMMA database access is publicly available at http://www.emmanet.org

133

COMANdrus : Community Annotation of the Chondrus crispus genome

Betina Porcel[1], Jonas Collén[2], Franck Aniere[1], Sylvain Bonneval[1], Benjamin Noel[1], Jean-Marc Aury[1], Corinne Da Silva[1], France Denoeud[1], François Artiguenave[1], Claude Scarpelli[1], Jean Weissenbach[1], Catherine Boyen[2], Patrick Wincker[1]

[1] Genoscope, Institut de Génomique, CEA, France
[2] UMR 7139 Marine Plants and Biomolecules, CNRS / UPMC-University of Paris VI, France

Presenter: Betina Porcel

Red seaweeds are key components of coastal ecosystems, economically important as food and source of gelling agents. Yet, they represent one of the last groups of complex multicellular organisms lacking a high-quality reference genome sequence. Annotation is one of the most difficult tasks in genome sequencing projects, yet it is essential for connecting genome sequence to biology. Most of the features that characterize a genome can be identified by the automated procedure used for the annotation. Although of very good quality, gene-modeling still remains tentative at the end of the automatic process, particularly when working with genomes with no sequenced close-related organism or not enough reliable resources available for automatic annotation. To attain both high throughput and quality data, a Genome Annotation workflow was set up at Genoscope for the florideophyte Chondrus crispus, based on automated annotation, community-wide genome analysis and expert validation to manually improve the annotated genome sequence. The Chon drus COM munity AN notation (COMANDrus) efforts were made possible using a distributed annotated platform, allowing the evaluation of the annotation through a remote connection to a Chado database linked to a Apollo graphical interface. Moreover, the Chondrus genome portal could be queried using genome browsers, BioMart and BLAST/BLAT servers, and be downloaded by biocurators to perform local searches. COMANdrus has supported research contributions from an active research community widespread in 10 different countries, resulting on curation of more than 25% of the annotated genes. Genome Annotation Workshops and Jamborees have been introduced as an integral part of this project, providing training in genome sequencing, annotation and analysis to researchers with different expertise. The participation of the Chondrus scientific community allowed us to ameliorate the understanding of the organism biology, as it will be presented.

134

A curation manual for the annotation of apoptosis in the Gene Ontology

Pablo Porras Millan, Paola Roncaglia, Rebecca Foulger, Jane Lomax, Rachael Huntley, Emily Dimmer, Sandra Orchard

EMBL-EBI, United Kingdom

Presenter: Pablo Porras Millan

The comprehensive and detailed analysis and depiction of biological pathways and reactions requires the use of databanks and ontologies in which previous knowledge is stored in an organized and standardized way and made available for the scientific community. Resources such as the Gene Ontology (GO, www.geneontology.org), a controlled vocabulary of terms describing gene product's characteristics, help to provide researchers with a reference from which they can infer current scientific knowledge, expanding and annotating their own datasets. The APO-SYS Consortium (www.apo-sys.eu/) is a project funded by the European Community to focus on "Apoptosis Systems Biology Applied to Cancer and AIDS", a thriving and fundamental field of biological research. Members of the consortium highlighted the need for a thorough revision of apoptosis-related terms in GO in order to properly depict current knowledge in the field. In a collaborative effort, members of the IntAct database and GO curators teams have undertaken a guided curation effort of literature recommended by the consortium experts, creating more granular terms that can accurately represent the scientific consensus in apoptosis. The next step would be to broaden the use of the updated terms in order to improve the annotation of gene products currently associated with less descriptive terms. In order to facilitate such effort, we have produced a specific set of curation guidelines for the apoptosis field of GO. This curation manual is primarily aimed to be used by GO curators and it describes the most relevant changes following the ontology structure from the parent term "cell death", outlining a decision tree that can be used to discern confusing or difficult annotation cases. The manual will be a useful companion to the re-structured ontology and that it can be of use beyond the scope of GO, providing any interested researcher or curator with a deeper understanding of the most up-to-date terms related with apoptosis.

135

Manual biocuration in UniProtKB/Swiss-Prot

Sylvain Poux1, Michele Magrane2, UniProt Consortium3

1 SIB Swiss Institute of Bioinformatics, Switzerland
2 EMBL-EBI, United Kingdom
3 UniProt Consortium, Switzerland

Presenter: Sylvain Poux

Manual curation of proteins is essential to provide high quality datasets that address the need of the scientific community for rapid access to reliable information which can in addition be programmatically parsed. Manually curated entries in the UniProt Knowledgebase (UniProtKB) are stored in the UniProtKB/Swiss-Prot section and constitute the highest priority of the UniProt Consortium, with more than 60% of its staff being fully dedicated to the task of biocuration High quality information is added by experienced biocurators, most of them with a strong background in research. The curation process involves extraction of a wide range of pertinent information from the scientific literature as well as manual verification of results from selected sequence analysis tools. Priority is given to the biocuration of proteins with an impact on the largest number of users. Information is reviewed, compiled, summarized and reported in the appropriate fields of a UniProtKB entry. UniProtKB/Swiss-Prot entries combine the manually verified protein sequence with experimental evidence derived from biochemical and genetic analyses, 3D-structures, mutagenesis experiments, information about protein interactions and post-translational modifications. We also manually assign Gene Ontology (GO) terms to all UniProtKB entries during the biocuration process based on experimental data from the literature. Quality has always constituted one of the highest priorities for UniProtKB/Swiss-Prot and it is essential to prevent the addition of erroneous data during manual curation and propagation of errors to automatically annotated records and to ensure the quality of imported data from external resources. We have developed a number of quality standards and procedures to ensure that the user community continues to have access to a high standard of data quality. This constant effort and the experience accumulated over the years have created a unique task force in the domain of manual biocuration.

136

User centered design in UniProt

Sangya Pundir

EMBL-EBI, United Kingdom

Presenter: Sangya Pundir

User centered design (UCD) is a product design approach that bases the process around information and feedback from the people who will use the product. UCD processes focus on users through the planning, design and development of a product. The UniProt user centered design process began with reviewing the current website with external users in June 2011. Findings from usability testing and log statistics informed an ongoing redesign of the UniProt web interface. We are following the UCD process by testing our designs with users from early stages, first using paper prototypes and then an interactive prototype site. Upon testing the prototype site with users from various backgrounds and usage levels, from occasional to expert users, we have seen a marked improvement in issues observed in usability tests in the 2011 website reviews. This process has helped us identify our user groups and their specific needs from UniProt. The implementation of this design is currently underway, with refinements based on feedback from niche groups, including curators. We are now focusing on presenting the UniProt sequence feature information in an interactive visual manner. The sequence feature viewer design has tested positively with users and is now under development. UniProt curators participated in an online card sorting game to help rearrange sequence features into suitable headings for the purpose of displaying them in the feature viewer. Also as part of this process, we aim to achieve consistent visualisations across difference resources at the European Bioinformatics Institute sharing common protein sequence features.

137

The challenge of increasing Pfam coverage of the human proteome

Jaina Mistry[1], Penelope Coggill[1], Ruth Eberhardt[1], Antonio Deiana[2], Andrea Giansanti[2], Robert Finn[3], Alex Bateman[1], Marco Punta[1]

[1] EMBL-EBI, United Kingdom
[2] Sapienza University of Rome, Italy
[3] HHMI, United States of America

Presenter: Marco Punta

It is a worthy goal to completely characterise all human proteins in terms of their domains. Here, we ask how far we have got in this endeavour. Ninety percent of proteins in the human proteome match at least one of 5,494 manually curated Pfam-A families. In contrast, human residue coverage by Pfam-A families is less than 45%, with 9,418 automatically generated Pfam-B families adding a further 10%. Even after excluding predicted signal peptide regions and short regions (<50 consecutive residues) unlikely to harbour new families, for about 38% of the human protein residues, distributed over almost 25,000 separate fragments, there is no information in Pfam about conservation and evolutionary relationship with other protein regions. Comparison to other sequences in the UniProtKB database suggests that the human fragments that exhibit similarity to thousands of other protein regions are often either divergent elements or N- or C-terminal extensions of existing families. 34% of fragments, on the other hand, match less than 100 regions in UniProtKB and do not overlap with existing Pfam-A families, which suggests that thousands of new families would need to be generated in order to cover them. Also, these latter fragments are particularly rich in regions of amino acid compositional bias, such as those predicted to be intrinsically disordered. This adds further complexity to the task of building them into new Pfam families. Based on these observations, our strategy for improving Pfam coverage of the human proteome in the near future will be a combination of improving the definition of existing families, and building new families whereas these have been experimentally functionally characterised.

138

MalaCards: an integrated compendium for diseases and their annotation

Noa Rappaport[1], Noam Nativ[1], Gil Stelzer[1], Michal Twik[1], Yaron Guan-Golan[2], Tsippi Iny Stein[1], Iris Bahir[1], Frida Belinky[1], Paul Morrey[3], Marilyn Safran[1], Doron Lancet[1]

[1] Weizmann Institute of Science, Israel
[2] LifeMap Sciences Inc., Hong Kong
[3] Utah Valley University, United Kingdom

Presenter: Noa Rappaport

Comprehensive disease classification, integration and annotation are crucial for biomedical discovery. At present, disease compilation is incomplete, heterogeneous and often lacking systematic inquiry mechanisms. We introduce MalaCards, an integrated database of human maladies and their annotations, modeled on the architecture and strategy of the GeneCards database of human genes. MalaCards mines and merges 44 data sources to generate a computerized card for each of 16,919 human diseases. Each MalaCard contains disease-specific prioritized annotations, as well as inter-disease connections, empowered by the GeneCards relational database, its searches, and GeneDecks set-analyses. First, we generate a disease list from 15 ranked sources, using disease-name unification heuristics. Next, we employ four schemes to populate MalaCards sections: 1) Directly interrogating disease resources, to establish integrated disease names, synonyms, summaries, drugs/therapeutics, clinical features, genetic tests, and anatomical context; 2) Searching GeneCards for related publications, and for associated genes with corresponding relevance scores; 3) Analyzing disease-associated gene-sets in GeneDecks to yield affiliated pathways, phenotypes, compounds, and GO terms, sorted by a composite relevance score and presented with GeneCards links; 4) Searching within MalaCards itself, e.g. for additional related diseases and anatomical context. The latter forms the basis for the construction of a disease network, based on shared MalaCards annotations, embodying associations based on etiology, clinical features and clinical conditions. This broadly disposed network has a power-law degree distribution, implying inherent properties of such networks. Work in progress includes hierarchical malady classification, ontological mapping, and disease set analyses, striving to make MalaCards an even more effective tool for biomedical research.

Database URL: http://www.malacards.org/

139

Improving gene-protein-reaction associations results in the enrichment of InterPro, GO and Rhea databases: the snowball effect of targeted manual curation

Claudia Rato da Silva, Alex Mitchell, Alessandro Vullo, Amaia Sangrador-Vegas, Siew-Yit Yong, Hsin-Yu Chang, Sarah Hunter, Jane Lomax, Paul Kersey

EMBL-EBI, United Kingdom

Presenter: Claudia Rato da Silva

Microme (http://www.microme.eu) is a resource for bacterial metabolism that aims at supporting the large scale inference of pathways directly from genome sequence. To facilitate this process, we developed a Genome-Reaction Matrix (GRM) containing inferred reactions from thousands of genomes, which can then be used to build draft metabolic networks and models. Reactions are defined in Rhea and ChEBI (resources of biochemical reactions and chemical entities) and their presence is inferred through the identification of genes encoding the enzymes responsible for their catalysis. This is done through a number of approaches, including protein classification using InterPro (a resource of predictive protein signatures) and the automatic functional annotation of classified proteins using Gene Ontology (GO) terms. In order to improve the accuracy and coverage of the inference methodology used in the GRM, we reviewed 1,211 InterPro entries related to transporters and 2,621 entries related to proteins involved in metabolism, by extracting evidence from published experimental data. Of these, 420 were associated with new or improved (i.e. more specific) GO terms. All GO Molecular Function terms assigned that could be translated into a Rhea reaction were cross-referenced, resulting in 113 new GO-Rhea mappings. During this process, new GO terms, reactions and chemical entities were created in the appropriate source databases as necessary. Overall, we achieved our main goal of increasing the number of gene-protein-reaction associations in the GRM: new annotations affected the reaction set of the vast majority of the genomes under analysis (99.6%; total of 4,149), increasing the number of gene-reactions associations by 8.3% (total of 3,722,698). This curation effort shows how targeting one database – InterPro – for enrichment can have positive knock-on effects on other resources and a cumulative snowball effect on annotation.

140

Identification and prioritization of novel, uncharacterized peptidases for biochemical characterization

Neil Rawlings

EMBL-EBI, United Kingdom

Presenter: Neil Rawlings

Genome sequencing projects are generating enormous amounts of biological data that require analysis, which in turn identifies genes and proteins that require characterization. Enzymes that act on proteins are especially difficult to characterize because of the time required to distinguish one from another. This is particularly true of peptidases, the enzymes that activate, inactivate and degrade proteins. This paper aims to identify clusters of sequences each of which represents the species variants of a single, putative peptidase that is widely distributed and is thus merits biochemical characterization. The MEROPS database maintains large collections of sequences, references, substrate cleavage positions and inhibitor interactions of peptidases and their homologues. MEROPS also maintains a hierarchical classification of peptidase homologues, in which sequences are clustered as species variants of a single peptidase; homologous sequences are assembled into a family; and families are clustered into a clan. For each family an alignment and a phylogenetic tree are generated. By assigning an identifier to a peptidase that has been biochemically characterized from a particular species (called a holotype) the identifier can be automatically extended to sequences from other species which cluster with the holotype. This permits transference of annotation from the holotype to other members of the cluster. By extending this concept to all peptidase homologues (including those of unknown function that have not been characterized) from model organisms representing all the major divisions of cellular life, clusters of sequences representing putative peptidases can also be identified. The 42 most widely distributed of these putative peptidases have been identified and discussed here and are prioritized as ideal candidates for biochemical characterization.

141

An HMM-based manual curation protocol for microbial protein families related to bioenergy

Fernanda Rego, Lucas M. Taniguti, Claudia B. Monteiro-Vitorello, João C. Setubal

University of São Paulo, Brazil

Presenter: Fernanda Rego

Hidden Markov Models (HMMs) are essential tools for automated annotation of protein sequences. For many years now protein family resources based on HMMs have been made available to the scientific community (e.g. TIGRfams). Much effort has also been devoted to the automated generation of protein family HMMs (e.g Panther). However, manually curated protein family HMMs remain the gold standard for use in genome annotation. Here we present a protocol for manual curation of protein family HMMs applied to microbial protein families of bioenergy interest. As input we assume a query sequence coding for a protein whose function has been experimentally demonstrated, linked to a publication, and manually annotated with Gene Ontology terms. We then apply the following steps: (1) definition of additional family member selection criteria: sequence size, minimum %-identity threshold to query sequence, minimum alignment coverage, presence of signal peptide or trans-membrane domain, and presence of conserved regions and domains; (2) search of additional family members on UniprotKB with BLAST. We use UniprotKB (TrEMBL + SwissProt) because its entries have richer records than those from GenBank; (3) multiple alignment and HMM building; (4) analysis of results and iteration of the process, with preliminary HMMs used for further searches; (5) definition of the minimum score threshold for the final HMM, or depending on family complexity, the establishment of subfamilies. Literature and web resource scans are also part of the process, so that we can link our results to existing related resources and assign comments and specific features to families or individual family members. We will present a case study of Glycoside hydrolase family 9 using as seeds proteins of the anaerobic species of Clostridium.

142

Utilising 3D Web Apps at eMouseAtlas

Lorna Richardson, Chris Armit, Shanmugasundaram Venkataraman, Peter Stevenson, Nick Burton, Julie Moss, Liz Graham, Yiya Yang, Jianguo Rao, Bill Hill, Richard Baldock

University of Edinburgh, United Kingdom

Presenter: Lorna Richardson

The eMouseAtlas resource has a wealth of 3D volumetric data, including embryo models with defined anatomical regions. Previously, it was necessary to use standalone applications to allow users full navigation through 3D data sets and to use them for spatial query. We have extended the standard Internet Imaging Protocol (IIP), developed for tile-based viewing of large scale 2D images, to 3D (IIP3D), which allows interactive visualisation of sections through 3D image volumes. IIP3D delivers tiled section views through very large 3D image volumes providing a capability not previously possible. In addition it enables viewing of multiple spatial domains overlaid on these sections. Our viewer is a fully interactive 3D web interface, allowing the user to navigate through the 3D space of the models and their associated spatial data. In addition, we are creating tools to 'paint' a 3D region and use this as the basis of a database query. We describe here the use of IIP3D viewer/tool within the context of eMouseAtlas and its additional application to other 3D atlas resources. The IIP3D client viewer is web-browser based and has been enhanced to use WebGL thereby delivering interactive 3D visualisation of the grey level data in the 3D context of the embryo.

143

Representation of apoptosis in the Gene Ontology

Paola Roncaglia, Pablo Porras Millan, Rebecca Foulger, Jane Lomax, Rachael Huntley, Emily Dimmer

EMBL-EBI, United Kingdom

Presenter: Paola Roncaglia

The Gene Ontology (GO) project aims at consistent descriptions of gene products across species and databases. To this purpose, GO provides terms to describe biological processes, molecular functions and cellular locations. As of Jan. 2013, GO contains >38,000 terms, used to create nearly 1.3 million high-quality manual annotations to gene products from many species. GO annotations are invaluable tools to the scientific community because they aid representation and interpretation of biological data, especially from high-throughput experiments. Recently, the APO-SYS Consortium, funded by the EC to focus on “Apoptosis Systems Biology Applied to Cancer and AIDS” and of which EBI is a member, highlighted the need for a thorough revision of apoptosis-related terms in GO. Apoptosis, also known as programmed cell death, is relevant during development and in various diseases such as cancer and disorders of the cardiovascular, neurological and autoimmune systems. The Apoptosis GO project was started and the ontology has been revised and enriched, so that current GO terms reflect the most up-to-date knowledge about apoptosis and other types of programmed cell death such as necroptosis. The ontology work is being paralleled by an annotation effort to increase number and information-content of annotations made to apoptosis-related proteins. The ultimate goal is to present the scientific community with a powerful resource, that can be used to highlight cell death mechanisms altered in some conditions (e.g. human cancer or drug treatment), and hopefully aid in discovering specific cell functions involved in apoptosis and cancer. Participation of apoptosis experts in this project has been fundamental; members of the APO-SYS consortium have been heavily involved in the set-up of the main guidelines and remain a reference point when complex issues emerge from the literature. The IntAct and Reactome databases have also taken part in this effort.

144

Uncovering hidden duplicated content in public transcriptomics data

Marta Rosikiewicz[1], Aurélie Comte[1], Anne Niknejad[2], Marc Robinson-Rechavi[1], Frederic Bastian[1]

[1] Swiss Institute of Bioinformatics - University of Lausanne, Switzerland
[2] University of Lausanne, Switzerland

Presenter: Marta Rosikiewicz

As part of the development of the database Bgee (a dataBase for Gene Expression Evolution), we annotate and analyze expression data from different types and different sources, notably Affymetrix data from GEO and ArrayExpress, and RNA-Seq data from SRA. During our quality control procedure, we have identified duplicated content in GEO and ArrayExpress, affecting about 14% of our data: fully or partially duplicated experiments from independent data submissions, Affymetrix chips re-used in several experiments, or re-used within an experiment. We present here the procedure that we have established to filter such duplicates from Affymetrix data, and our procedure to identify future potential duplicates in RNA-Seq data. Database URL: http://bgee.unil.ch/

145

Average rank IQR – a new improved method for Affymetrix microarray quality control for metaanalysis and database curation

Marta Rosikiewicz, Marc Robinson-Rechavi

University of Lausanne/SIB, Switzerland

Presenter: Marta Rosikiewicz

Over the past few years thousands of microarray results have become available for exploratory and metaanalysis studies. The quality control step that allows elimination of poor quality arrays is essential for developing valuable databases dedicated to these tasks. The classical quality assessment methods that are intended for identification of outlier arrays within a single experiment may be insufficient to eliminate experiments where the majority of samples are of low quality, thus the development of new methods is necessary. In the current study we tested known methods for quality assessment of Affymetrix microarrays along with two new methods proposed by us – average rank Inter Quantile Range (arIQR) and PM/MM t-test. As an independent measure of quality we specified how well the gene expression profile from each array correlates with reference expression profile of homologous genes in the same organ from different species. We discovered that samples, which show lower correlation with the reference, could be identified as of poor quality on the basis of some of the quality metrics. The newly proposed method arIQR outperforms all the other methods for this task and is implemented in Bgee (http://bgee.unil.ch), the database built for evolutionary comparison of gene expression patterns between animal species.

146

Raw Read Data in ENA: Tools and Methods

Marc Rossello

EMBL-EBI, United Kingdom

Presenter: Marc Rossello

The European Nucleotide Archive (ENA) at the European Bioinformatics Institute (EBI) has constituted Europe's primary nucleotide sequence resource for nearly 30 years. Traditionally curating, archiving and presenting assembled sequence and annotations, the ENA has broadened its services in recent years to encompass the raw output from next generation sequencing technologies. This high volume data type has merited a new model for sequence archival and metadata capture to maximise downstream interpretation and analysis by the community. By archiving raw machine output we provide extensive capacity to preserve, analyse and compare the nature of samples from around the globe. In support of this, we operate an extensive set of minimal reporting standards developed in house and strive to collect MIxS (Minimal Information about a Sequence) attributes to describe environmental samples, MIGS (Minimal Information about a Genome Sequence) attributes to describe isolated organism samples for genome sequencing and MINSEQE (Minimum Information about a high-throughput SeQuencing Experiment) attributes for expression and/or differentiation experiments. Here we discuss the tools and methods available for data submission and data discovery with regards to the raw read data in ENA. The focus is on an uncomplicated automated system with an interactive web interface and an accompanying help desk contactable by email and staffed by ENA biocurators and other appropriate team members. The data model divides metadata (run, sample, experiment, and analysis objects) from the data themselves. All aspects of a next generation sequencing study can be effectively described and subsequently browsed with these interconnecting objects. Checklists embedded into the metadata capture steps enable us capture standards-compliant data.

147

The miRandola Database: Function and Diagnostic Potential of Extracellular microRNAs.

Francesco Russo[1], Sebastiano Di Bella[1], Giovanni Nigita[1], Valentina Macca[1], Alessandro Laganà[2], Rosalba Giugno[1], Alfredo Pulvirenti[1], Alfredo Ferro[1]

[1] University of Catania, Italy
[2] The Ohio State University, United States of America

Presenter: Francesco Russo

MicroRNAs are small noncoding RNAs that play an important role in the regulation of various biological processes through their interaction with cellular mRNAs. They have shown great potential as tissue-based markers for cancer classification and prognostication. miRNAs are also present in extracellular human body fluids such as serum and plasma. Since miRNAs circulate in the bloodstream in a highly stable form, they may be used as blood-based biomarkers for many diseases. miRandola is a comprehensive manually curated classification of extracellular miRNAs. The current version of the database contains 2132 entries, with 581 unique mature miRNAs and 21 types of samples. miRNAs are classified into four categories, based on their extracellular form: Ago2, exosome, HDL and circulating. The latter is used when authors do not distinguish the extracellular form and constitutes the largest group. To date, miRandola contains information gathered from 89 papers. miRandola is connected to miRò, the miRNA knowledge base, allowing users to infer the potential biological functions of miRNAs and their connections with phenotypes. miRandola is an ongoing project. Here we describe the new features which are currently being developed: (i) new data gathered from more than 110 papers, (ii) information on different species (e.g. Mus musculus, Rattus norvegicus) and different extracellular miRNA forms (e.g. Microvesicles), (iii) new sources are currently being added to the system (e.g. Vesiclepedia, a compendium for extracellular vesicles), (iv) integration of the DAVID Web Service that will allow users to get a functional annotation chart report and a functional annotation clustering report to infer the potential biological function of extracellular miRNAs through enriched biological themes, particularly GO terms and Kegg pathways, and (v) integration of the BioMart framework to export tables in text or Excel format. miRandola is available online at: http://atlas.dmi.unict.it/mirandola/.

148

The MetaboLights repository: curation challenges in metabolomics

Reza Salek

EMBL-EBI, United Kingdom

Presenter: Reza Salek

MetaboLights is the first general-purpose, open-access curated repository for metabolomic studies, their raw experimental data and associated metadata, maintained by one of the major open-access data providers in molecular biology. Increases in the number of depositions, number of samples per study and the file size of data submitted to MetaboLights present a challenge for the objective of ensuring high quality and standardised data in the context of diverse metabolomic workflows and data representations. Here, we describe the MetaboLights curation pipeline, its challenges and its practical application in quality control of complex data depositions. Database URL: http://www.ebi.ac.uk/metabolights

149

InterPro2Go: manual curation as a starting point for automatic GO term annotation

Amaia Sangrador-Vegas, Hsin-Yu Chang, Siew-Yit Yong, Alex Mitchell, Sarah Hunter

EBI-EMBL, United Kingdom

Presenter: Amaia Sangrador-Vegas

About 98% of all GO annotations in UniProtKB are inferred through automated annotation systems. The InterPro2GO pipeline is one of the main sources of such annotations for uncharacterised proteins. It provides annotation to over 60% of sequences in UniProtKB, assigning tens of millions of individual GO terms. Underlying the InterPro2GO pipeline is the manual assignment of GO terms to InterPro entries by curators. This process depends on the existence of experimental evidence for characterised sequences from published papers. Based on biological knowledge, curators decide which functions, processes and components, and at what level of granularity, can be propagated to the sets of proteins that match an InterPro entry. In this poster we explain briefly how GO terms are assigned to InterPro entries, as well as showing the extent of coverage of InterPro2GO for different proteomes.

150

BioSharing and bioDBcore: establishing the lay of the standards landscape and databases

Philippe Rocca-Serra[1], Eamonn Maguire[1], Alejandra Gonzalez-Beltran[1], Pascale Gaudet[2], Susanna-Assunta Sansone[1]

[1] University of Oxford e-Research Centre, United Kingdom
[2] SIB Swiss Institute of Bioinformatics, Switzerland

Presenter: Susanna-Assunta Sansone

Community standardization initiatives work to develop minimum information checklists, terminologies and file formats, which are increasingly used in the structuring, description and curation of data sets. These standards aim to ensure that descriptions of entities of interest (e.g., receptors) and related assays contain sufficient contextual information (e.g., provenance of materials, technology and measurement types) to be comprehensible, in principle reproducible and reusable; without such context, data are of little value. The BioSharing works to catalogue available community standards [http://www.biosharing.org], extending the work started with the Minimum Information for Biological and Biomedical Investigations (MIBBI) portal [http://mibbi.sf.net] and linking to existing entries in the BioPortal [http://bioportal.bioontology.org]. BioSharing intends to promote community standards which already exist, discouraging redundant if unintentional competition, and work as a registry for new efforts. BioSharing allies with the International Society for Biocuration (ISB) and in collaboration with NAR Database issue works to catalogue databases, described according to the BioDBcore guideline [Gaudet, et al. NAR Database, 2011]. These metadata descriptors also require information about the minimum information checklist(s), terminology(s) and file format(s) used by the database to represent and serve its datasets. This allows linking databases with relevant community standards, providing information on their usage and value. By establishing the lay of the standards landscape, in time it will also be possible to (i) build graphs of relationships and complementarities in scope and functionality and (ii) progressively link with several other existing resources' portals and catalogues, contributing to the distributed ecosystem of interconnected resources, valuable to curators, researchers, funders and publishers.

151

Automated Mouse Phenotype Annotations for IMPC

Luis Santos[1], Julian Atienza-Herrero[1], Andy Blake[2], Chao-Kung Chen[3], Armida Di Fenza[1], Richard Easty[4], Simon Greenaway[1], Alan Horne[4], Natasha Karp[4], Vivek Yver[4], Gautier Koscielny[3], Jeremy Mason[3], Terrence Meehan[3], David Melvin[4], Hugh Morgan[5], Asfand Qazi[4], Ahmad Retha[1], Duncan Sneddon[1], Jonathan Warren[3], Henrik Westerberg[1], Robert Wilson[4], Gagarine Yaikhom[1], Steve Brown[1], Paul Flicek[3], Helen Parkinson[3], William Skarnes[4], Ann-Marie Mallon[1]

[1] MRC Harwell, United Kingdom
[2] Medical Research Council, United Kingdom
[3] EMBL-EBI, United Kingdom
[4] Wellcome Trust Sanger Institute, United Kingdom
[5] MRC Mammalian Genetics Unit, United Kingdom

Presenter: Luis Santos

The International Mouse Phenotyping Consortium (IMPC, www.mousephenotype.org) is an international effort to phenotype up to 20,000 mutant mouse lines during the next decade. This will generate knockout mouse strains from mouse embryonic stem cells for each protein-coding gene. Gathering the efforts of phenotyping centres worldwide and informatics from MRC Harwell, European Bioinformatics Institute, and the Wellcome Trust Sanger Institute (Mouse Phenotyping Informatics Infrastructure, MPI2), IMPC has established a high-throughput phenotyping pipeline (www.mousephenotype.org/impress) assessing different aspects of mouse biology. As with any high-throughput system, a considerable amount of data will be produced, requiring new and intelligent solutions to be put in place before new scientific knowledge can be effectively elucidated. A major component of these intelligent solutions includes automated classification, curation and annotation, currently being implemented by MPI2. MPI2's main task is to find and characterise the phenotypic effect of each of the gene knockouts on the mutant lines, based on statistical comparison to the wildtype mice. Previous experience in a similar, smaller scale project (www.eumodic.org) provided us with knowledge and experience to devise tools capable of dealing with the diversity, variability and nature of IMPC's data. This includes simple Boolean data, 2 and 3D images, numerical data, different measurement units and free text. Such variety presents various challenges, adding to inherent difficulties like validating the data and verifying it is within the limits of biological plausibility but without obscuring real phenotypic effects. The automated annotation process we are developing will allow rapid release of the data and results and also makes for an efficient use of human resources, so we may also focus on overcoming future challenges, such as integrating this data within existing sources and discovering new links to human disease.

152

Structure Integration with Function, Taxonomy and Sequence (SIFTS)

Sanchayita Sen

EMBl-EBI, United Kingdom

Presenter: Sanchayita Sen

In recent decades we have witnesses an explosion in the volume of biological data being generated. This has necessitated the development of resources to archive, annotate, distribute and manage those data. Linking and integrating the information captured across these data resources has now become vital to facilitate the knowledge-discovery process and drive forward biomedical research. The challenge of maintaining up-to-date cross-reference information for macromolecular structures is addressed by the Structure Integration with Function, Taxonomy and Sequences resource (SIFTS; http://pdbe.org/sifts). SIFTS is an on-going collaboration between the Protein Data Bank in Europe (PDBe, http://pdbe.org) and UniProt (http://www.uniprot.org/). The two teams, based at the European Bioinformatics Institute (EBI), have since 2002 maintained a semi-automated process for providing up-to-date residue level mappings between UniProt and PDB entries. SIFTS also provides cross-references between PDB structures and other biological resources such as Pfam, SCOP, CATH, GO, InterPro and the NCBI taxonomy database. In this talk we will describe how SIFTS data are widely used to support bioinformatics resources and underpin many tools for accessing and analysing structural data. We will also demonstrate how the up-to-date annotation data in SIFTS makes it possible for non-expert users to locate structural information using familiar biological terms and classification systems such as genes, proteins, pathways, enzyme nomenclature, sequence-family information (Pfam) and GO annotations.

153

The Guide to PHARMACOLOGY: expert-driven curation of pharmacological targets and the substances that act on them

Joanna Sharman[1], Adam Pawson[2], Helen Benson[2], Elena Faccenda[2], Stephen Alexander[3], John Peters[4], Alistair Mathie[5], John McGrath[6], Anthony Davenport[7], Michael Spedding[8], Anthony Harmar[2], NC-IUPHAR[9]

[1] University of Edinburgh, United Kingdom
[2] The University/BHF Centre for Cardiovascular Science, United Kingdom
[3] School of Biomedical Sciences, United Kingdom
[4] Neuroscience Division, United Kingdom
[5] Medway School of Pharmacy, United Kingdom
[6] School of Life Sciences, United Kingdom
[7] Clinical Pharmacology Unit, United Kingdom
[8] Les laboratoires Servier, France
[9] The International Union of Basic and Clinical Pharmacology Committee on Receptor Nomenclature and Drug Classification, United Kingdom

Presenter: Joanna Sharman

The Guide to PHARMACOLOGY portal (http://www.guidetopharmacology.org) is being developed as a joint initiative between the British Pharmacological Society (BPS) and the International Union of Basic and Clinical Pharmacology (IUPHAR), with the aim of providing an open access resource covering all aspects of pharmacology. The first version links together two authoritative, expert-driven resources, together covering >2200 established, or potential, human drug targets and >5000 substances that act on them. Information on drug targets comes from the BPS Guide to Receptors and Channels (GRAC), published as a free supplement to the British Journal of Pharmacology, and now integrated in a relational database with data from IUPHAR-DB (http://www.iuphar-db.org). GRAC provides expert overviews of the key pharmacological properties of G protein-coupled receptors, ion channels, nuclear hormone receptors, catalytic receptors, transporters and enzymes, with the key licensed medicines and experimental drugs that act on them, and recommended reading lists for newcomers to each field. For selected targets, IUPHAR-DB provides expanded peer-reviewed information on the pharmacological, genetic, functional and pathophysiological characteristics as well as comprehensive lists of agonists, antagonists and modulators with quantitative data from the literature. Links are provided to corresponding entries in other relevant databases and to citations in PubMed. Drugs, endogenous substances and radioligands are annotated with manually curated 2D chemical structures, calculated physico-chemical properties, IUPAC name and synonyms. The Guide to PHARMACOLOGY ultimately intends to cover all of the human targets of current prescription medicines and other likely targets of future small molecule drugs, and to become an authoritative global resource intelligible to all members of the scientific community, in order to maximise our expanding knowledge of how druggable genes affect health and disease.

154

Cross-Species Functional Analysis of Gene Sets: The Gene Annotator

Mary Shimoyama, Jeff de Pons, RGD Team

Rat Genome Database, United States of America

Presenter: Mary Shimoyama

Analysis of gene sets is commonly part of genome, microarray, GWAS and SNP analysis studies and is a time consuming endeavor. Analyzing large sets of genes or even single genes across species is cumbersome and requires multiple steps. The Rat Genome Database Gene Annotator Tool is a one stop functional analysis tool for rat, human and mouse genes. Users can upload a list of common identifiers from RGD, EntrezGene, GenBank, Ensembl and Affymetrix or search by a chromosomal region or functional ontology identifiers to retrieve comprehensive 3 species functional reports for disease and phenotype, pathway, Gene Ontology, and drug/chemical-gene interactions as well as dozens of identifiers and links to other sources. The Genome Plot provides a genome view of positions for genes in the set, map data, the ability to overlay other data such as QTLs and provides direct links to GBrowse where users can add other tracks such as SNP, QTL, disease or transcripts. Functional analysis of the entire gene list or subsets can be accomplished through the Annotation Distribution Tool and the Comparison Heat Map. The Annotation Distribution Tool provides a dynamic assessment of the functional make-up of the gene list, showing the percentage of genes associated with various diseases, pathways, biological processes and functions. Users retrieve genes associated with a particular disease, pathway or function and further analyze this subset for functional commonalities. The Comparison Heat Map visualizes the distribution of genes across two functional parameters such as disease and pathway and users can drill down to more specific categories or change functional categories. Genes in each block are displayed and annotation distribution analysis available for each subset. Gene Annotator is currently used to analyze genes from multiple studies including patient sequencing projects.

155

Using Phenomics to Identify Mouse Models Of Age-Related Human Disease

Michelle Simon, Saumya Kumar, Simon Greenaway, Siddharth Sethi, Karen Pickford, Laura Wisby, Andy Blake, Paul Potter, Ann-Marie Mallon

Medical Research Council, United Kingdom

Presenter: Michelle Simon

MRC Harwell conducts a number of large scale high throughput initiatives to deduce genotype to phenotype associations in mutant mice. One such program is the Harwell Aging Screen. This phenotype-driven screen mutagenizes male mice (G0) with the chemical N-ethyl-N-nitrosourea (ENU) which induces random point mutations in the spermatogonia; subsequent generations (G3) of mutant mice are then subjected to an array of phenotype procedures and monitored for phenodeviants. The screen plans to generate approx. 200 G3 pedigrees, capture their phenotypes over an 18 month period and identify lines with age related phenotypes, ranging from neurological to metabolic to behavioural phenotypes. We have designed a database to hold 18 months' worth of temporal phenotype data for each mutant line. Each phenotype parameter is attributed a mammalian ontology term adapted from the EmpressSlim pipeline used previously. Pedigrees with an array of interesting phenotypes will have their G1 grandparent subjected to whole genome next-generation sequencing (WGS) in order to find the exhaustive list of sequence variants (SNPs, small INDELs and structural variants) which may contribute to phenodeviants in the G3 pedigree. All novel mutations in the WGS require stringent analysis to associate the genuine ENU-induced or spontaneous mutation(s) with the phenotype. Similarly temporal phenotype data and associated MPs require novel statistical methods to identify significant groups of phenodeviants. Here we present our efforts to associate the phenotypic groups to possible causative mutations by comparing results obtained from robust high throughput pipelines and novel statistical methods. These results will provide a deeper understanding of the processes leading from genomic changes to age related diseases

156

Phyletic Profiling with Cliques of Orthologs is enhanced by Signatures of Paralogy Relationships

Nives Škunca[1], Matko Bosnjak[2], Anita Krisko[3], Pance Panov[4], Saso Dzeroski[4], Tomislav Smuc[2], Fran Supek[2]

[1] ETH, Switzerland
[2] Rudjer Boskovic Institute, Croatia (Hrvatska)
[3] MedILS, Croatia (Hrvatska)
[4] Jozef Stefan Institute, Slovenia

Presenter: Nives Škunca

New microbial genomes are sequenced at a high pace, allowing insight into the genetics of not only cultured microbes, but a wide range of metagenomic collections such as the human microbiome. To understand the deluge of genomic data we face, computational approaches for gene functional annotation are invaluable. We introduce a novel model for computational annotation that refines two established concepts: annotation based on homology and annotation based on phyletic profiling. The phyletic profiling-based model that includes both inferred orthologs and paralogs—homologs separated by a speciation and a duplication event, respectively—provides more annotations at the same average Precision than the model that includes only inferred orthologs. For experimental validation, we selected 38 poorly annotated Escherichia coli genes for which the model assigned one of three GO terms with high confidence: involvement in DNA repair, protein translation, or cell wall synthesis. Results of antibiotic stress survival assays on E. coli knockout mutants showed high agreement with our model's estimates of accuracy: out of 38 predictions obtained at the reported Precision of 60%, we confirmed 25 predictions, indicating that our confidence estimates can be used to make informed decisions on experimental validation. Our work will contribute to making experimental validation of computational predictions more approachable, both in cost and time. Our predictions for 998 prokaryotic genomes include ~400000 specific annotations with the estimated Precision of 90%, ~19000 of which are highly specific—e.g. "penicillin binding," "tRNA aminoacylation for protein translation," or "pathogenesis"—and are freely available at http://gorbi.irb.hr/

157

The Genomic Standards Consortium

Peter Sterk

Oxford e-Research Centre, University of Oxford, United Kingdom

Presenter: Peter Sterk

The Genomic Standards Consortium (GSC) [1] was established in 2005 to tackle the challenge of working towards better descriptions of genomes and metagenomes through community-level, consensus-driven solutions. The GSC's mission is to work towards i) the implementation of new genomic standards, ii) methods of capturing and exchanging the information captured in these standards (metadata, or contextual data) and iii) harmonization of information collection and analysis efforts across the wider genomics community. Thus far, the GSC has created a standard, the Minimum Information about any (x) Sequence (MIxS), which includes three minimum information checklists for describing genomes, metagenomes, and environmental marker sequences (MIGS/MIMS/MIMARKS) upon submission to the public databases and publication. MIxS requires core information on habitat, geolocation, and sequencing methodology as well as fields specific to data type and a range of optional environmental packages to capture core measurements defining a broad range of habitats, including water, soil, and host-associated habitats. Large, well-contextualized genome, metagenome, and marker gene data sets (e.g., ribosomal gene surveys) provide ideal opportunities for comparison and contrasting using computational means to solve a wide range of questions in biology. The GSC is becoming into a hub for the coordination of large-scale projects. For any standard to create a lasting impact requires substantial input from the wider scientific community, including adoption and support. The GSC urges researchers interested in pushing the boundaries of genomic science through collaboration to join and contribute expertise to building the GSC roadmap for the future. 1. Field D et al. (2011) The Genomic Standards Consortium. PLoS Biol 9(6): e1001088. doi:10.1371/journal.pbio.1001088

158

MicroB3 community Standards

Petra ten Hoopen[1], Guy Cochrane[1], MicroB3 Consortium[2]

[1] EMBL-EBI, United Kingdom
[2] http://www.microb3.eu/partners, United Kingdom

Presenter: Petra ten Hoopen

Standards-compliance becomes a crucial requirement for any dataset aiming to be recognised and discoverable by its research community and beyond. The MicroB3 community standards are being developed by the MicroB3 Consortium that aims to create a platform for smooth handling and analysis of marine microbial sequence data in their environmental context. The MicroB3 standards-compliant sample metadata shall be i/ harmonised with the standardised description of (meta)genomes and marker genes promoted by the INSDC (http://www.insdc.org/) and GSC MIxS standards (http://gensc.org/gc_wiki/index.php/MIxS), ii/ compliant with minimal reporting requirements for oceanographic data supported by the SeaDataNet marine profile of the ISO19139 content standard for describing geographic datasets (http://www.seadatanet.org/) and iii/ compliant with the OBIS extension of the Darwin Core standard reference for describing biodiversity information (http://www.iobis.org/). The MicroB3 standards will consist of minimal reporting requirements and standard operating procedures covering a sampling site, sample and data processing. All components of the MicroB3 standards will be described in the Ocean Sampling Handbook and available to participants of the Ocean Sampling Day, a simultaneous sampling campaign of the world's oceans to reveal a marine microbial diversity. Advanced standards-compliance of marine microbial sample data and technical advances to handle it are prerequisites for data integration across research domains. However, a dedicated effort of curators at oceanographic, biodiversity and bioinformatics data centres will be necessary to support the databases' interoperability. Apart from fostering standard protocols and checklists of their local resource curators shall validate accuracy of core classifiers carrying provenance information that will connect data across resources and shall re-direct inappropriately routed data submissions to the relevant integrated archive.

159

Annotation of genes with Gene Ontology terms in an evolutionary context

Rama Balakrishnan[1], Pascale Gaudet[2], James Hu[3], Eva Huala[1], Ranjana Kishore[4], Suzanna Lewis[5], Donghui Li[1], Brenley McIntosh[3], Huaiyu Mi[6], Li Ni[7], Paul Thomas[6]

[1] Stanford University, United States of America
[2] SIB Swiss Institute of Bioinformatics, Switzerland
[3] Texas A&M University, United States of America
[4] California Institute of Technology, United States of America
[5] Lawrence Berkeley National Laboratory, United States of America
[6] University of Southern California, United States of America
[7] The Jackson Laboratory, United States of America

Presenter: Paul Thomas

The Gene Ontology (GO) has been used widely to annotate the functions of gene products, based on published experimental results obtained in a number of different “model” organisms. The GO Consortium has implemented a project to integrate and review annotations made for related genes in different organisms, and to use this information to improve both the quality and quantity of GO annotations. Quality is addressed through simultaneous manual review of experimental annotations for related genes in different organisms, leading to removal of erroneous or misleading annotations and increased annotation consistency. Quantity is addressed through use of experimental annotations from a few model organisms to infer annotations for related genes in many different organisms. The basic approach has been published (Gaudet et al., Briefings in Bioinformatics 12:449, 2011), and involves reviewing the annotations in the context of a phylogenetic tree of a gene family and then creating a model of function evolution within the family. We will discuss our progress and lessons learned, as well as challenges and implications for further development of both curation tools and automated approaches to large-scale functional annotation.

160

European Nucleotide Archive (ENA) Advanced Search

Ana Toribio, European Nucleotide Archive

EMBL-EBI, United Kingdom

Presenter: Ana Toribio

The European Nucleotide Archive launched, in November 2012, an advanced search tool with the goal of improving discoverability of data and enabling their interactive exploration in real time. The complexity of the sequencing data stored (with more than 250 million annotated and assembled sequences and more than 200 million annotated features), the daily update cycle, and the need to provide query response times (typically below 10 seconds) represent a significant challenge, especially for complex queries. To power ENA's advanced search, we have established a new data warehouse, built upon analytical database technology, that supports extremely large numbers of objects, hundreds of queriable attributes, geospatial queries and simplicity of configuration and use. We will present this advanced search tool and provide a user perspective on the first of our interfaces to the tool, an interactive query builder. In addition, further plans of improvement such as supporting additional indexed domains, indexed fields, and additional functions like geospatial polygons will be presented. Finally, we will outline future more intuitive interfaces that are in development.

161

Manual Biocuration in the European Nucleotide Archive (ENA): Still Necessary?

Ana Toribio, European Nucleotide Archive

EMBL-EBI, United Kingdom

Presenter: Ana Toribio

A core activity of the ENA is the biocuration of the assembled and annotated sequences. The biocuration process involves the understanding of the sequence nature and the integration of all relevant biological information into the database in an accurate and comprehensive way. Moreover, the easy access to public data that ENA provides as a basis for computational analysis is based on the appropriate annotation of the sequences. It is the manual biocuration effort that ensures that both experimental and inferential results are annotated correctly. A particular case is the taxonomic classification of the organism to which the nucleotide sequence belongs to. Taxonomy curation represents not only a key step in the manual biocuration process but also an example of the INSDC collaboration. Manual biocuration is also the pillar to update the database records as new information becomes available. In response to the flood of nucleotide sequence data, ENA has introduced a check list submission system that simplifies not only the submission route but also the biocuration process. This can be seen as an automatic curation tool. However, the complexity and the diversity of the information do not allow its implementation for all potential submissions. Furthermore, when the system does not cover the full understanding of a particular sequence, this is finally left to criteria that are managed by curators. Manual biocuration involves not only a critical and deep review of the submitted sequence but also –if necessary- a thorough conversation between the curator and the submitter in order to achieve the established quality of the full annotation. The nature of this conversation is –as our records show- one of the main components of the submitter user experience, highlighting the level of communication required in the curation process. Hence, manual biocuration is still a crucial strategic activity to maintain the standards of the European Nucleotide Archive.

162

Nematode Variation in WormBase

Mary Ann Tuli, Kevin Howe

EMBl-EBI, United Kingdom

Presenter: Mary Ann Tuli

WormBase is the canonical repository for information on C.elegans and closely related nematodes. We have recently made changes in how variation data is stored and presented in response to both large volumes of data from Whole Genome Sequencing (WGS) projects, and community feedback. A significant change has been to distinguish between naturally-occurring polymorphisms and laboratory-induced alleles. Laboratory-induced alleles are identified by the allele designation of the laboratory of origin, whereas naturally-occurring polymorphisms are identified by the WormBase variationID. We have also begun the process of merging SNPs from different strains at identical location and with the same base change. Previously, a SNP which had been identified in multiple naturally-occurring strains would have been represented as separate variation objects for each strain. Such SNPs will now be regarded as a single reference variation in which we list all the strains which carry the SNP. We currently have variation data from 9 WGS projects. Four are from mutagenised strain projects, while the other 5 are from wild isolate strain projects. In total, these projects have generated nearly 1.2 million new variations. A new version of the WormBase website was launched in March 2012 with significant improvements to the way variation data can be viewed by users. Coloured fields in the Strain widget on the Variation Summary Page allow users to very easily see which strains carry a variation and whether the strain is available from the CGC. The Gene Summary Page now allows users to choose the way Variations are viewed. It is now possible to sort Variations by a various criteria, including type of molecular change, effect on the protein, and the number of associated phenotypes. We have also increased the complement of variation tracks on the genome browser, providing the users with finer-grained control over which subsets of variations are displayed.

163

Representing Drosophila models of human disease in FlyBase

Susan Tweedie, Gillian Millburn, FlyBase-Cambridge Curators, Nicholas H. Brown

University of Cambridge, United Kingdom

Presenter: Susan Tweedie

Drosophila is used extensively to model a wide variety of human diseases. In an effort to promote the utility of flies in medical research and to help researchers to find relevant genetic resources, FlyBase has started to curate models of human disease using the Disease Ontology (http://disease-ontology.org/). We define a model of disease as any modification of Drosophila resulting in a phenotype that recapitulates some aspect of human disease. Papers that contain descriptions of new disease models or utilize existing disease models are flagged during our literature triage process. At present our disease curation is limited to models with a genetic basis but we are tracking other types of model (e.g. chemically or environmentally induced) with a view to future curation. We have designed our disease annotations to be similar in structure to Gene Ontology (GO) annotations, both to make it easier for users, already familiar with GO, to interpret the disease annotations and to make it simpler for literature curators who will be making both types of annotation. Another advantage of this approach is that it allows us to reuse existing syntax-checking and data-loading code from our curation pipeline and avoid major schema changes. The disease annotations are found in a new section of the gene report devoted to the relationship between human and fly genes; this section includes orthology information and links to resources such as HUGO gene nomenclature committee (HGNC) and OMIM. This poster will present our annotation model and summarize our disease curation progress.

164
Curating the Microbiology literature for an Ontology of Microbial Phenotypes

Peter Uetz[1], Marcus Chibucos[2], Deborah Siegele[3], Michelle Giglio[2], James Hu[3]

[1] Peter Uetz, United States of America
[2] University of Maryland School of Medicine, United States of America
[3] Texas A&M University, United States of America

Presenter: Peter Uetz

Bacteria have only few easily visible morphological phenotypes such as colony shape, so mutants have to be characterized using physiological characters or other features. Nevertheless, mutant phenotypes are critical to make inferences about the functions of genes and the organization of gene products into pathways and processes. Phenotypes are also the basis of classifying and identifying microbes. Comparing phenotypes across systems can provide insight into the similarities and differences in their underlying biology, and suggest new experiments. To that end, the Ontology of Microbial Phenotypes (OMP) captures standardized phenotypic information from bacteria and other microbes. OMP terms capture different levels of specificity through specified relationships, and have English (reader friendly) definitions and synonyms to facilitate connecting formal terms to the unstandardized literature. Wherever possible, OMP will build on other ontologies, including the Gene Ontology (GO) and Phenotypic Quality Ontology (PATO). OMP is developing the formal ontology as well as gathering examples of phenotype statements from the literature that must be expressible in the ontology. Currently, our curation efforts focus on the Escherichia coli literature, including both large-scale genetic screens as well as detailed studies of individual genes. We have established a wiki for annotation and for browsing the ontology. We anticipate that diverse user groups will employ OMP for annotation and analysis of microbial phenotypes, similar to GO. Definitions of phenotypes and links to the original literature should facilitate the use of phenotype data to understand functions and relationships of genes, cells, and systems, including interacting organisms such as bacteria and their phage.

http://microbialphenotypes.org/wiki/index.php/Main_Page

165

Curating taxonomic information from 250 years of scientific literature

Peter Uetz

Peter Uetz, United States of America

Presenter: Peter Uetz

Taxonomy is the foundation of all biology, providing a system of species and higher taxa. Although >100 taxonomic databases and several meta-databases exist (such as the Catalogue of Life [CoL] and the Encyclopedia of Life [EoL]), there is still no complete list of all species on the planet. The problem is exacerbated by the fact that the taxonomic literature lasts back more than 250 years, starting with Linné's Species Plantarum in 1753. Thus, all taxonomic databases have to curate literature that may have been published in the 18th century. We illustrate the process using the Reptile Database, a database of nearly 10,000 species, and put it in relation to the CoL and EoL, both of which rely on this and other species databases for their taxonomies. All taxonomic databases face multiple issues: the continuously changing names and ranges of organisms, leading to synonyms, the inconsistent definition of species, and the integration of other information, such as DNA sequences, images, maps, multimedia, or phenotypes, and their links to other data sources such as GenBank or Wikispecies. Col, EoL, and the Biodiversity Heritage Library (BHL) cover only a small fraction of the taxonomically relevant literature, given that the BHL focuses on material that is out of copyright. More efficient and automated curation will be required in the future.

http://www.reptile-database.org
http://www.catalogueoflife.org/
http://www.eol.org/

166

Mining the Phenotype-Disease Associations from the Scientific Literature

Drashtti Vasant[1], Jane Lomax[1], Terrence Meehan[1], Paul Schofield[2], Damian Smedley[3], Maria Taboada[4], Helen Parkinson[1], Peter Robinson[5]

[1] EMBL-EBI, United Kingdom
[2] University of Cambridge, United Kingdom
[3] Wellcome Trust Sanger Institute, United Kingdom
[4] University of Santiago, Spain
[5] Charite-Univeritatsmedizin Berlin, Germany

Presenter: Drashtti Vasant

Advances in translational research and analyses of complex disease present a challenge for data integration as patient data with associated clinical information relating to common and complex diseases are now a common research tool e.g. Diabetes. This presents a challenge for integration with existing data, as although there are many resources describing disease in different contexts the disease-phenotype relationship is often implicit. The human phenotype ontology provides a structured description for human phenotype, and these can be linked to disease resources e.g. OMIM. However, the coverage for common phenotype and explicit links to disease for these is incomplete, and inaccessible in a query-able resource. Here we describe an experiment to mine human phenotype-disease associations from the biomedical literature for four disease areas Marfan syndrome, cardiomyopathy, diabetes type 2 and breast cancer using the CiteXplore web service and available abstracts. Our hypothesis is that by mining the literature the frequency of co-occurring terms from specific dictionaries will provide a set of automatically literature mined disease-phenotype associations. A pilot experiment with Marfan syndrome, and five text mining dictionaries (Human-phenotype ontology, Mammalian phenotype ontology, Mouse pathology, Chemical entities of biological interest (CheBI) and Diseases (UMLS)) and curation of results as gold standard set by an expert clinician indicated good recall, but suboptimal precision. Subsequent mining with a reduced set of dictionaries is expected to improve precision, for example, removing differential diagnoses, and further processing of abstracts to include negation, and targeting conclusions is expected to improve these further.

167

Annotating the Biomedical Literature for the Human Variome

Karin Verspoor[1], Antonio Jimeno-Yepes[1], Lawrence Cavedon[1], Tara McIntosh[2], Asha Herten-Crabb[3], Zoë Thomas[3], John-Paul Plazzer[4]

[1] National ICT Australia, Australia
[2] Wavii, Inc., United States of America
[3] The University of Melbourne, Australia
[4] Royal Melbourne Hospital, Australia

Presenter: Karin Verspoor

We introduce the Variome Annotation Schema, a schema that aims to capture the core concepts and relationships relevant to cataloguing and interpreting human genetic variation and its relationship to disease, as described in the published literature. The schema was inspired by the needs of the database curators at the International Society for Gastrointestinal Hereditary Tumours (InSiGHT) database, but is intended to have application to genetic variation in a range of diseases. The schema has been applied to a small corpus of full text journal publications on the subject of inherited colorectal cancer. The schema and the corpus represent a unique resource for the development of text mining solutions that address relationships among patient cohorts, disease, and genetic variation.

168

Documents classification system for the kinase-specific biomedical literature

Dina Vishnyakova

University and University Hospitals of Geneva, Switzerland

Presenter: Dina Vishnyakova

We present the original integration of an automatic text categorization pipeline that we developed to perform classification and prioritization functions in order to speed up the biocuration of the kinase-specific biomedical literature. Biocuration is a complex task, it requires domain expertise and specific training. The task performed by a professional curator can be simplified into the following workflow: Retrieval of documents given a particular query (e.g. name of kinase) in a particular document repository; Selection of articles: the selection of a biocurator is usually based on the title and the abstract; Reading of a particular article where the full-text article is read by the biologist; Extraction of information: a particular passage is analyzed to obtain a representation of the level of entities such as proteins, diseases, gene ontologies; Normalization of the extracted information, when the biologist transforms a particular passage into a normalized identifier or a set of normalized descriptors. The system we designed covers all steps. The task of the system can be basically described as a classification task, where the system presents three scoring functions to rank a selected set of articles by its relevance to defined domains of interest: gene ontologies, protein-protein interaction and pathologies. Then components of a question-answering system are used to extract kinase-specific annotations from the ranked list of articles. Each ranking is based on a ruled-based scoring function which combines three main modules: an information retrieval engine for MEDLINE (EAGLi), a gene normalization (GN) service (NormaGene) developed for a BioCreative campaign and finally, a set of answering components and entity recognizer for diseases and gene ontologies. The main components of the pipeline are publically available both as web application and web services. The specific integration is available via a web user interface at http://pingu.unige.ch:8080/Kinase .

169

A Taxonomy for Immunologists

Randi Vita, James Overton, Jason Greenbaum, Alessandro Sette, Bjoern Peters

La Jolla Institute for Allergy & Immunology, United States of America

Presenter: Randi Vita

The NCBI Taxonomy of organisms is a widely known, utilized, and respected resource. To facilitate standardization and interoperability, the IEDB has utilized NCBI Taxonomy for annotation of host organisms and epitope sources since its inception. However, navigation through the many levels of the taxonomy can be daunting for immunologists. For example, 'Homo sapiens' is located 27 levels deep in the NCBI tree. Although the IEDB search interface showed only those branches of the NCBI tree that were used by the IEDB and provided text search, the tree was often overwhelming for users. We recently developed techniques for extracting subsets of the NCBI Taxonomy into simpler, shallower trees for specific purposes. We provide a minimal but familiar level of parentage, and add synonyms commonly used by immunologists, while retaining the benefits of the original ontology, including standardized nomenclature, synonyms, identifiers, and relationships. Using multiple trees we provide different views of the taxonomy, including a full view of all taxa used in IEDB, and a partial view of only the organisms that are sources of allergens. Our simplified taxonomic trees are easy to maintain and navigate. For example, we reduced the depth of 'Homo sapiens' to 5 levels, with each level familiar to any immunologist. We believe our process is informative and can benefit the users of other databases presenting organism related data.

170

Ontology analyses of disease objects at The Rat Genome Database

Shur-Jen Wang, Stanley Laulederkind, Mary Shimoyama, G. Thomas Hayman, Jennifer Smith, Timothy Lowry, Victoria Petri, Rajni Nigam, Melinda Dwinell

Medical College of Wiscon, United States of America

Presenter: Shur-Jen Wang

The Rat Genome Database is the premier resource for genetic, genomic and phenotype data for the laboratory rat, Rattus norvegicus. The RGD team focuses on manual curation of gene-disease associations for rat, human and mouse. In this work, we have analyzed disease-associated strains, quantitative trait loci (QTL) and genes from rats. Among disease portals, the Cardiovascular Disease and Obesity /Metabolic Syndrome Portals have the highest number of rat strains and QTL. These two portals share 398 rat QTL, and these shared QTL are highly concentrated on rat chromosomes 1 and 2. For disease-associated genes, we performed gene ontology (GO) enrichment analysis across portals using RatMine enrichment widgets. Two BP terms, 'blood vessel development' and 'regulation of programmed cell death' were the most highly enriched BP terms across all portals except in the Obesity/Metabolic Syndrome Portal where 'lipid metabolic process' was the most highly enriched BP term. 'Cytosol' and 'nucleus' were very common Cellular Component annotations for disease genes in the portals, but only the Cancer Portal genes were highly enriched with 'nucleus' annotations. Similar enrichment patterns were observed in a parallel analysis using the DAVID Functional Annotation Tool. The relationship between GO terms and disease terms was examined reciprocally by retrieving rat genes annotated with the GO terms found enriched in the disease portals. The individual GO term annotated gene list shows enrichment in physiologically-related diseases. For example, the 'regulation of blood pressure' genes are enriched with cardiovascular disease annotations, and the 'lipid metabolic process' genes with obesity annotations. Furthermore, we were able to enhance enrichment of neurological diseases by combining 'G-protein coupled receptor binding' annotated genes with 'protein kinase binding' annotated genes.

171

Annotation of the Expression Atlas using the Experimental Factor Ontology

Eleanor Williams, Tony Burdett, Simon Jupp, James Malone, Benedetto Fiorelli, Maria Keays, Nataliya Kryvych, Jan Taubert, Misha Kapushesky, Robert Petryszak, Helen Parkinson, Alvis Brazma

EMBL-EBI, United Kingdom

Presenter: Eleanor Williams

The Expression Atlas (www.ebi.ac.uk/gxa) is a semantically enriched database of gene expression data. Searches can be made to find up and down regulated genes in particular conditions (e.g. diabetes or after treatment with a chemical compound) or to find conditions in which a particular gene is differentially expressed. The sample properties and experimental factors (variables) in each study are annotated against an application ontology called the Experimental Factor Ontology (EFO, www.ebi.ac.uk/EFO) which is designed for annotation of functional genomics data. The ontology is used in the Atlas search interface to expand queries of experimental factor conditions to include child terms, allow querying of synonyms and to suggest query terms. For example, searching for 'cancer' returns matches to 'carcinoma' and 'leukemia' in addition to 'cancer'. Annotation of sample properties and experimental factors to EFO is achieved by both automated and manual methods. Zooma (www.ebi.ac.uk/fgpt/sw/zooma/) uses a knowledgebase of curator reviewed annotations to automatically generate optimal sample and factor annotations for all new studies added to the Atlas. Corona software (www.ebi.ac.uk/fgpt/sw/corona/) is used for manual curation of new and complex terms. It uses NCBO BioPortal and Annotator services to suggest annotations from EFO and can be used to add or update annotations. The EFO mappings also facilitate the linking of samples and treatments to other databases. Links with ChEBI ('small' chemical compounds) already exist and plans are underway to integrate gene expression, proteomics and metabolomics data as well as domain-specific expression databases such as the Mouse Expression Spatial Database (EMAGE). EFO will be used to annotate the new Baseline Expression Atlas to be released in 2013. This resource will provide information about which gene products are present in tissues and cell types in several species.

172

Sequencing and Comparative Analysis of the Gorilla MHC Genomic Sequence

Laurens Wilming[1], Elizabeth Hart[1], Penelope Coggill[2], Roger Horton[1], James Gilbert[1], Chris Clee[1], Matt Jones[1], Christine Lloyd[1], Sophie Palmer[2], Sarah Sims[1], Siobhan Whitehead[1], David Wiley[1], Stephan Beck[3], Jennifer Harrow[1]

[1] Wellcome Trust Sanger Institute, United Kingdom
[2] EMBL-EBI, United Kingdom
[3] University College London, United Kingdom

Presenter: Laurens Wilming

Major histocompatibility complex (MHC) genes play a critical role in vertebrate immune response and because the MHC is linked to a significant number of auto-immune and other diseases it is of great medical interest. Here we describe the clone-based sequencing and subsequent annotation of the MHC region of the gorilla genome. Because the MHC is subject to extensive variation, both structural and sequence-wise, it is not readily amenable to study in whole genome shotgun sequence such as the recently published gorilla genome. The variation of the MHC also makes it of evolutionary interest and therefore we analyse the sequence in the context of human and chimpanzee. In our comparisons with human and re-annotated chimpanzee MHC sequence we find that gorilla has a trimodular RCCX cluster, versus the reference human bimodular cluster, and additional copies of Class I (pseudo)genes between Gogo-K and Gogo-A (the orthologues of HLA-K and HLA-A). We also find that Gogo-H (and Patr-H) is coding versus the HLA-H pseudogene and, conversely, there is a Gogo-DQB2 pseudogene versus an HLA-DQB2 coding gene. Our analysis, which is freely available through the VEGA genome browser, provides the research community with a comprehensive dataset for comparative and evolutionary research of the MHC.

173

UbiGRID: Gene annotation, modification sites and interactions in the ubiquitin-proteasome system.

Andrew Winter[1], Julie Nixon[1], Andrew Chatr-Aryamontri[2], Lorrie Boucher[3], Bobby-Joe Breitkreutz[3], Christie Chang[4], Daici Chen[2], Sven Heinicke[4], Nadine Kolas[3], Michael Livstone4, Lara O'Donnell[3], Rose Oughtred[4], Lindsay Ramage[1], Teresa Reguly[3], Jennifer Rust[4], Chris Stark[3], Kara Dolinski[4], Mike Tyers[2]

[1] Wellcome Trust Centre for Cell Biology, University of Edinburgh, United Kingdom
[2] Institute for Research in Immunology and Cancer, Université de Montréal, Canada
[3] Systems Biology, Samuel Lunenfeld Research Institute, University of Toronto, Canada
[4] Lewis-Sigler Institute for Integrative Genomics, Princeton University, United States of America

Presenter: Andrew Winter

The ubiquitin-proteasome system (UPS) regulates many cellular processes by the covalent attachment of ubiquitin to substrate proteins. Ubiquitination proceeds by activation of ubiquitin by an E1 enzyme, conjugation to an E2 enzyme and then ligation to lysine residues in substrates or their growing polyubiquitin chains via interaction of substrate and E2 with an E3 enzyme. The fate of ubiquitinated proteins is determined by a variety of ubiquitin binding domains, which can direct the substrate for degradation by the 26S proteasome, or alter substrate localization, interactions or activity. Deubiquitinating enzymes can act in opposition to E3s by removing ubiquitin or shortening polyubiquitin chains. Specificity in the UPS is dictated in large part by E3 enzyme-substrate interactions; notably the human genome encodes over 800 potential E3 enzymes, many of which can target multiple substrates. In an attempt to identify and curate the multitude of substrates and enzymes of the UPS we have created the UbiGRID resource within the BioGRID database (http://www.thebiogrid.org). It includes: 1) annotation for ~1200 human and ~270 budding yeast genes likely to be involved in the UPS, divided into 22 UPS-related functional categories, e.g. E3, E2, DUB. 2) 110,302 interaction evidences representing 60,595 unique pairings of 15,017 genes curated from 10,252 papers, with maximal coverage of the UPS achieved by targeting PubMed searches to all UPS gene names. 3) 47,862 ubiquitination sites in 11,919 substrate proteins from 15 high throughput mass spectrometry studies. Analysis of the UPS networks reveal that the UPS genes connect to ~35% of the human and yeast genomes. A dedicated UbiGRID project web page (http://ubigrid.thebiogrid.org) allows viewing and download of UPS gene annotation, interactions and ubiquitination sites. Work is now ongoing to use the interaction data to generate hypotheses for UPS biological functions and involvement in disease.

174

Publications from a Biocurator's Point of View

Ulrike Wittig, Renate Kania

Heidelberg Institute for Theoretical Studies, Germany

Presenter: Ulrike Wittig

The main task of our (biocurators) daily work is to extract information from publications to populate databases and enrich database content. For further usage the information has to be stored in a structured and standardized format. From exchange of experiences with other biocurators we know, that we are not alone with the problem to be faced with incomplete, missing or ambiguous data every now and then. This was the reason to decide for a closer look at publications from a biocurator's point of view regarding the structure and content of papers. Therefore we checked articles published during the last 50 years. These were randomly selected from papers used for data extraction for the SABIO-RK database. SABIO-RK (http://sabio.h-its.org) is a web-accessible database containing biochemical reactions and their kinetic properties. Beside the information related to kinetic properties we also investigated general data like detailed information about proteins catalyzing the biochemical reactions. As a special focus we analyzed for example the frequency of the usage of external database identifiers. Since SABIO-RK not only stores information about reactions and their kinetic properties but also experimental conditions under which kinetic parameters were measured we also had a closer look at the correctness and completeness of the assay conditions. Additionally we investigated the distribution of information within the articles and checked for different formats in which data are represented in the publications. Based on our findings we will recommend guidelines for authors and publishers to improve the reusability of their papers.

175

The YeastGenome App: the Saccharomyces Genome Database at your fingertips

Edith Wong[1], Kalpana Karra[1], Benjamin Hitz[2], Eurie Hong[1], J. Michael Cherry[1]

[1] Stanford University, United States of America
[2] SGD, United States of America

Presenter: Edith Wong

The Saccharomyces Genome Database (SGD) is a scientific database that provides researchers with high-quality curated data about the genes and gene products of Saccharomyces cerevisiae. To provide instant and easy access to this information on mobile devices, we have developed YeastGenome, a native application for the Apple iPhone and iPad. YeastGenome can be used to quickly find basic information about S. cerevisiae genes and chromosomal features regardless of internet connectivity. With or without network access, you can view basic information and Gene Ontology annotations about a gene of interest by searching gene names and gene descriptions or by browsing the database within the app to find the gene of interest. With internet access, the app provides more detailed information about the gene, including mutant phenotypes, references, and protein and genetic interactions, as well as provides hyperlinks to retrieve detailed information by showing SGD pages and views of the genome browser. SGD provides online help describing basic ways to navigate the mobile version of SGD, highlights key features, and answers frequently asked questions related to the app. The app is available from iTunes (http://itunes.com/apps/yeastgenome). The YeastGenome app is provided freely as a service to our community, as part of SGD’s mission to provide free and open access to all its data and annotations.

Database URL: http://www.yeastgenome.org

176

Using biological process co-annotation for ontology and annotation quality control

Valerie Wood[1], Antonia Lock[2], Midori Harris[1], David Hill[3], Harold Drabkin[3], Maria Costanzo[4], Seth Carbon[5], Chris Mungall[5]

[1] Cambridge University, United Kingdom
[2] University College London, United Kingdom
[3] Mouse Genome Informatics, United Kingdom
[4] Stanford University, United States of America
[5] Lawrence Berkeley Institute, United States of America

Presenter: Valerie Wood

BACKGROUND Biological processes are accomplished by the coordinated action of sets of gene products. However, some processes are rarely connected to each other because they functionally, temporally or spatially distant. We speculated that we could identify pairs of biological processes which were unlikely to be co-annotated to the same gene products (e.g. amino acid metabolism and cytokinesis), and use any mutually exclusive processes identified to create rules which alert curators to possible annotation errors. METHODS Co-annotated terms (annotation intersections) were identified for all pairs of "high level" GO processes for three taxonomically distant organisms (fission yeast, budding yeast, mouse). Intersections where annotations were "null" were used to create rules of the form "x is not usually co-annotated with y". Intersections where annotations were sparse were inspected for spurious annotations, which were either corrected, or, if the annotations were validated the rules were extended to allow these exceptions. RESULTS In the rule generation phase of this work, we have identified cases that account for accurate co-annotation as well as several types of error that give rise to incorrect co-annotation. We present descriptions and examples for each type of correct and incorrect case, and indicate how errors have been addressed. The rules will be incorporated into the central GO annotation quality control system, where they can be applied to the entire annotation corpus, allowing further refinement of the rule base as well as identifying, and reducing the occurrence of annotation errors.

177

PomBase.org

Valerie Wood[1], Mark McDowall[2], Midori Harris[1], Antonia Lock[3], Kim Rutherford[1]

[1] Cambridge University, United Kingdom
[2] EMBL-EBI, United Kingdom
[3] University College London, United Kingdom

Presenter: Mark McDowall

PomBase (http://www.pombase.org) is a new model organism database to support the organization of and access to scientific data for the fission yeast Schizosaccharomyces pombe. PomBase will support genomic sequence and features, genome-wide datasets and manual literature curation. Gene overview pages present the data related to a gene, including the gene type, product, sequence features, phenotypes, Gene Ontology annotation, modifications and physical and genetic interactions. The data is housed within an Ensembl genome database, and the gene overview pages link through to an interactive, Ensembl-style, genome browser. A query builder has been implemented to allows users to search by multiple feature types. A query history summarises queries and allows queries to be combined and edited. Results pages provide access to gene overview pages. The Ensembl-style genome browser, which is accessible from the gene overview pages, provides the functionality to store, analyse and visualise a wide variety of datasets mapped to the genome either from internal sources or via externally loaded URLs or data files. Examples of supported datasets that can be imported include whole genome resequencing data, ChIP-chip and ChIP-seq assay, mapping to microarray probes and other high-throughput data types. The Ensembl-style browser also provides views of orthologous and syntenic regions by comparative analysis of closely related genomes (S. japonicus and S. octosporus). PomBase will also provide a community hub for researchers, providing genome statistics, a community curation interface, news, events, documentation FAQs and mailing lists.

178

Let’s Talk About Sets

Matt Wright, Kris Gray, Ruth Seal, Elspeth Bruford

HUGO Gene Nomenclature Committee (HGNC), United Kingdom

Presenter: Matt Wright

The HUGO Gene Nomenclature Committee (HGNC) has assigned unique approved gene symbols and names to over 34,000 human loci to date. Over 19,000 of these are protein coding genes, but we also name pseudogenes, phenotypic loci, genomic features and non-coding RNAs. Our website, genenames.org, is a searchable repository of HGNC approved gene nomenclature and associated resources. Each locus has an individual “gene symbol report” which can include links to genomic, proteomic and phenotypic information. Approved gene symbols are based on names describing structure, function or homology, and where possible the HGNC also organise these into gene groupings and families, many of which have specialist advisors who are experts in that particular area of biology. HGNC also create web pages for specific groups of genes; these “gene sets” are mostly grouped together by homology and/or function but sometimes by other shared information such as structure or genomic location. Recently we have greatly expanded the scope of these pages, including hub pages listing all the constituent subclasses for some large gene families such as the G-protein-coupled receptors and zinc finger gene families. We also now provide links from our individual gene symbol reports to our improved gene family resources. If you know of a gene family that you think we should include or update, please contact us via hgnc@genenames.org or talk to us during this meeting.

179

A Curation Assistance System Using Natural Language Processing Technologies

Yasunori Yamamoto[1], Synobu Okamoto[1], Seiha Miyazawa[2], Natsuko Ichikawa[2], Shuji Yamazaki[2], Nobuyuki Fujita[2]

[1] Database Center for Life Science, Research Organization of Information and Systems, Japan
[2] Biological Resource Center, National Institute of Technology and Evaluation, Japan

Presenter: Yasunori Yamamoto, Synobu Okamoto

Recent development of sequencing technology brings out the massive flow of new genome sequence data. Scientists are required not only to register the genome sequences to International Nucleotide Sequence Databases (INSD), but also to annotate encoded genes in a comprehensive manner. Assigning protein definition is a major part and one of a rate-limiting step in the process of genome annotation. Especially, it is difficult to annotate genes in comprehensive manner with controlled vocabulary. This difficulty is mainly caused by referring various gene descriptions annotated separately with different standards and vocabulary. Development of more effective annotation system assigning protein descriptions in comprehensive manner will assist rapid submission of genome sequence and highly curated annotation. We use natural language processing (NLP) technology to absorb lexical variants and normalize them. There are several patterns to generate those variants, from simpler differences such as presence or absence of hyphen or singular or plural forms to more complicate ones such as synonyms of gene or protein names. Especially for the synonyms, we need to take a special care because we have to consider species or consistency of the entire annotations. The annotators usually make rephrasing rules explicitly or implicitly, so we work on codifying them to make the task computable. For example, two protein definitions are obtained from automatic annotation processes: cell cycle protein FtsW and cell division protein FtsW. Both are valid per se, but adjustment to the former is needed to follow the in-house rule. After obtaining a list of rules, we checked them to see if there are any inconsistencies or contradictions among them. Here we report the results and lessons learned, and discuss future works.

180

H-InvDB gene family/group: an annotation resource for all human gene families and groups

Chisato Yamasaki[1], Akiko Noda[2], Yuichiro Hara[2], Takashi Gojobori[3], Tadashi Imanishi[4]

[1] National Institute of Biomedical Innovation (NIBIO) and BIRC, AIST, Japan
[2] Biomedicinal Information Research Center (BIRC), National Institute of Advanced Industrial Science and Technology (AIST), Japan
[3] National Institute of Genetics, Japan
[4] Tokai University School of Medicine, Japan

Presenter: Chisato Yamasaki

Here we report the curation procedures and database description of curated human gene families and groups of integrated database of human genes, transcripts and proteins, H-Invitational Database (H-InvDB). Gene family/group view provides manually curated datasets for four selected human gene families or groups, Immunoglobulin (Ig), Major Histocompatibility Complex (MHC), T-cell receptor (TCR) and Olfactory receptor (OR). The automatic annotation was based on the genomic location, similarity against known proteins and gene name and symbols of the known gene groups, and then finally all the results were curated manually and double checked. Also, we provide the annotation data set of predicted human gene families. After excluding the four selected gene families/groups, clusters of duplicated genes were defined by the single linkage clustering and then family names were assigned manually based on the most common InterPro domains. We defined an unique identifier for each H-Invitational gene family/group (HIF). H-InvDB 8.0 provides annotation for 35,631 protein-coding human gene clusters based on human full-length cDNAs and mRNAs. Among then, we identified 3,313 H-Inv human gene families consisting of 17,892 H-Inv gene clusters, including 1,007 new human gene family candidates of unknown function. We provide these annotation data at Gene family/groups view (http://www.h-invitational.jp/hinv/genefamily/index_en.cgi) which consists of two main views, the index view for the overview and the detailed view for each annotation dataset. Hyperlinks to the related H-InvDB viewers and public databases; HGNC, InterPro and IMGT-LIGM were provided. We also provide useful data mining tools such as "Navigation search" that enables searches of gene families with free combination of 16 search contents, such as genomic location or alternative splicing patterns.

181

The InterPro web site - refreshed, revamped, refined

Siew-Yit Yong

EMBL-EBI, United Kingdom

Presenter: Siew-Yit Yong

InterPro, a resource of predictive protein signatures, was originally established more than 12 years ago. Whilst the database has grown steadily since its inception and additional analyses have been added, for the last 6 years the web site has remained largely static. Here we describe the recently relaunched InterPro web site, which aims to increase functionality and enhance the experience of our users. The new web site offers an improved results summarisation page, clearer display of entry relationships, a summary view for domains and sites, easier browsing across species, 1-click download of sets of sequences, as well as a clearer and cleaner interface. The layout of the new entry, results, species and individual protein pages are highlighted here.

182

Curation at the Protein Data Bank

Jasmine Young[1], Sanchayita Sen[2], Reiko Igarashi[3]

[1] RCSB PDB, United States of America
[2] EMBL-EBI, United Kingdom
[3] PDBj, Japan

Presenter: Jasmine Young

The Protein Data Bank (PDB) is the single archive for 3D macromolecular structures. The archive serves as a primary and critical resource for research in structural biology and in drug discovery worldwide. The quality of the data in the archive is regularly reviewed to support the community of PDB users who require consistent and highly accurate data for their scientific research. This presentation describes the processes and tools used by the Worldwide PDB (wwPDB) to maximize data quality of individual structures and across the archive. This includes the development of a new deposition and annotation system that will improve the quality and effectiveness of the data curation process.

The index number refers to the abstract number and not the page number.

184

Integration of the transcriptional regulation of carbon sources, in Escherichial coli K-12, with their central metabolism and other cellular systems

Peralta-Gil M., Ledezma-Tejeida D., Gama-Castro S., Santos-Zavaleta A. and Collado-Vides J.

UNAM, Mexico

Presenter: Daniela Ledezma-Tejeida

RegulonDB is a database that integrates mainly the biological knowledge of the mechanisms that regulate transcription initiation in Escherichia coli and, eventually, the knowledge on its complete regulatory network. Thus, we have followed different strategies in order to better describe the complexity of the biology within the database, such as curating by publication date, on the bases of regulons and sigmulons, by functional class, by biological system, by binding sites of a particular transcription factor (TF), including other regulatory mechanisms, and finally by elementary genetic sensory response unit, or Gensor unit (GU). In this sense, to attain an understanding of bacterial metabolic adaptations to carbon source availability, currently we have curated about 40 regulons related to carbon source utilization, and we have diagrammed 15 GUs involved in the uptake and degradation of different carbon sources. Our main goal of this specific integration is to facilitate understanding of E. coli metabolic adaptations to carbon source availability and the coupling between the different genetic regulatory layers within the central metabolism. We are integrating and modeling the information on the metabolic pathways for the different carbon sources with the available information at the molecular level, including the interplay between transcriptional regulation and regulatory mechanisms, such as allosteric interactions and the effects of secondary products. Thus, these interactions at the molecular level lead to adjustments in metabolism between glycolytic and gluconeogenic carbon sources and the efficient adaptation between the electron transport chain, different two-component systems, and the acid stress response, among other cellular systems. The unified view of transport, metabolism and gene regulation help to better understand both metabolism and the regulation adaptation to alternative carbon sources encoded in the genome of E.coli K-12.