Database, Vol. 2012, Article ID bas036, doi:10.1093/database/bas036

Original article

Recent advances in biocuration: Meeting Report from the fifth International Biocuration Conference

Pascale Gaudet[1,*], Cecilia Arighi[2], Frederic Bastian[3], Alex Bateman[4], Judith A. Blake[5], Michael J. Cherry[6], Peter D'Eustachio[7], Robert Finn[8], Michelle Giglio[9], Lynette Hirschman[10], Renate Kania[11], William Klimke[12], Maria Jesus Martin[13], Ilene Karsch-Mizrachi[12], Monica Munoz-Torres[14], Darren Natale[14], Claire O'Donovan[13], Francis Ouellette[15], Kim D. Pruitt[12], Marc Robinson-Rechavi[3], Susanna-Assunta Sansone[16], Paul Schofield[17], Granger Sutton[18], Kimberly Van Auken[19], Sona Vasudevan[14], Cathy Wu[2], Jasmine Young[20] and Raja Mazumder[21,*]

[1]Chair, International Society for Biocuration and CALIPHO Group, Swiss Institute of Bioinformatics, 1 Rue Michel Servet, Geneva, Switzerland, [2]University of Delaware, Newark, DE, USA, [3]Swiss Institute of Bioinformatics and Department of Ecology and Evolution, University of Lausanne, Lausanne, Switzerland, [4]Wellcome Trust Sanger Institute, The Wellcome Trust Genome Campus, Hinxton, UK, [5]The Jackson Laboratory, Bar Harbor, ME, [6]Department of Genetics, Stanford University, Stanford, CA, [7]NYU School of Medicine, New York, NY, [8]HHMI - Janelia Farm Research Campus, VA, [9]Institute for Genome Sciences, University of Maryland School of Medicine, Baltimore, MD, [10]The MITRE Corporation, Bedford, MA, USA, [11]Heidelberg Institute for Theoretical Studies, Heidelberg, Germany, [12]NCBI/NLM/NIH/DHHS, Bethesda, MD, USA, [13]European Bioinformatics Institute, Wellcome Trust Genome Campus, Hinxton, UK, [14]Georgetown University, Washington, DC, USA, [15]Ontario Institute for Cancer Research and Department of Cell and Systems Biology, University of Toronto, Toronto, ON, Canada, [16]University of Oxford, Oxford, UK, [17]Cambridge University, Cambridge, UK, [18]J. Craig Venter Institute, Rockville, MD,[19]California Institute of Technology, Pasadena, CA, [20]RCSB Protein Data Bank, Rutgers University, NJ, and [21]Chair, Fifth International Biocuration conference organizing committee; Department of Biochemistry and Molecular Biology, The George Washington University, Washington, DC 20037, USA

***Corresponding author:** Tel: +41 22 379 49 17; Fax: +41 22 379 58 58; Email: pascale.gaudet@isb-sib.ch

Correspondence may also be addressed to Raja Mazumder. Tel: +1 202 994 5951; Fax: +1 202 994 9994; Email: mazumder@gwu.edu

Submitted 4 July 2012; Revised 13 September 2012; Accepted 19 September 2012

The 5th International Biocuration Conference brought together over 300 scientists to exchange on their work, as well as discuss issues relevant to the International Society for Biocuration's (ISB) mission. Recurring themes this year included the creation and promotion of gold standards, the need for more ontologies, and more formal interactions with journals. The conference is an essential part of the ISB's goal to support exchanges among members of the biocuration community. Next year's conference will be held in Cambridge, UK, from 7 to 10 April 2013. In the meanwhile, the ISB website provides information about the society's activities (http://biocurator.org), as well as related events of interest.

Introduction

Biological databases have been a key item in the toolbox of life scientists for 30 years. Initially mostly focused on annotation of sequences, a wealth of highly varied resources have burgeoned in the past 10 years, propelled in part by the development of high-throughput techniques and the resulting acceleration in data generation. Those databases are developed and maintained by biologists and computer scientists who have the specific expertise of creating tools and platforms for indexing, integrating, displaying and ultimately helping understand complex biological data. There are a large number of biological databases, as indicated by the 2012 Nucleic Acids Research (NAR) online database collection that catalogued 1380 published biological databases (1). In this context, professional biocuration is now becoming well established. The International Society for Biocuration (ISB, http://www.biocurator.org) is a

Database, Vol. 2012, Article ID bas036, doi:10.1093/database/bas036

non-for-profit organization founded in 2009 to promote the interests of biocurators. The ISB now counts over 300 members from nearly 150 databases and institutions in 26 different countries. This corresponds to only a fraction of the biocuration community, as several groups such as biocurators from commercial databases, as well as researchers, students and post-docs who perform biocuration work as part of a research project are under-represented in the ISB.

Since 2005, the biocuration community has been holding conferences with the goals of presenting and promoting the various projects, exchanging ideas, and fostering collaborations among the biocuration community. Those conferences have been extremely successful, and their popularity continues to increase. The 5th International Biocuration Conference was held in Georgetown, Washington, DC, USA, from 2 to 4 April 2012. Hosted by The Protein Information Resource (PIR), the conference provided a great opportunity for attendees to interact with other groups interested in biocuration. Over 300 biocurators and researchers attended the conference, with delegates from over 100 universities, research institutes and companies and representing 17 different countries.

Conference highlights

The conference was organized around seven sessions and five workshops, summarized in the next sections of the article. Two poster sessions were held with over 70 posters at each event. A best poster prize was awarded to Nives Skunca for her work on assessing the quality of non-experimental curated and electronic Gene Ontology (GO) annotations. Professors Mark Yandell (Department of Human Genetics, University of Utah, USA), Frederick P. Roth (Donnelly Centre for Cellular and Biomolecular Research, University of Toronto, and Samuel Lunenfeld Research Institute, Mt. Sinai Hospital, Toronto, Canada) and Amos Bairoch (Department of Structural Biology and Bioinformatics, Faculty of Medicine, University of Geneva, Geneva, Switzerland) gave stimulating plenary lectures. Yandell described his work to develop tools for annotating genomes and their sequence-variants using interoperable, machine-readable data standards. Roth discussed technologies for mapping and navigating genomes and genetic networks. Bairoch gave an overview of his pioneering work in biocuration, from Swiss-Prot to his new group CALIPHO that has two missions for one goal: increasing our knowledge on human proteins via integration of information on human proteins in a new database, neXtProt, and through the experimental characterization of proteins of unknown function.

Community annotation

The conference started with a session on community annotations. The talks covered a diverse set of approaches for capturing biological annotations unified by the common goal of trying to engage scientists to help curate data. This is a topic of great interest to most databases, as well as their users, as a possible means to help increase efficiency of data capture.

One popular way of capturing annotations from the community is through the use of wikis. This approach was presented by Nicholas Stover, for the *Tetrahymena* genome database, as well as other ciliate genomes, namely *Ichthyophthirius multifiliis* and *Oxytricha trifallax* (1). Andrew Su presented the GeneWiki project, a Wikipedia-based annotation tool for human genes (2). One of the key points of using Wikipedia is the sheer number of editors who have produced, detailed and information-rich articles. Furthermore, Su described how the processing of the GeneWiki annotations suggests novel Gene Ontology (GO) and disease associations.

The Skate (*Leucoraja erinacea*) genome database, presented by Cathy Wu, has employed a different approach to community engagement, through workshops and annotation jamborees (3). This approach provided scientists a structured way to disseminate knowledge, thereby giving rise to community intelligence of the annotation process.

Representatives from FlyBase and PomBase presented solutions aimed at increasing the involvement of their research communities in the annotation process. Gillian Millburn described how FlyBase contacts authors of research papers to provide a synopsis of their paper's content via a simple web-based form. This facilitates prioritization of papers for detailed manual biocuration by FlyBase curators. Antonia Lock presented PomBase's web-based CANTO tool that allows the coupling of papers, genes and annotations using controlled vocabularies. The tool is already being used internally by the database curators and will shortly be released to the *Schizosaccharomyces pombe* research community.

From these talks, it is clear that given the appropriate tools, the wider scientific community can be involved in a more distributed annotation model. Databases should harness the researchers' willingness to provide information by creating simple, yet robust mechanisms for contributing biological annotations. Community annotation needs to be complemented by the work of database biocurators to ensure consistency and quality, as well as to expand the areas where community annotations are incomplete and design new tools and data models as new techniques are developed.

Functional annotation and pathways

Pathway databases associate an organism's proteins with molecular functions, represent these as reactions, and group the reactions based on shared components: the output of one reaction might be the input, catalyst or regulator of a second, and so on. Alexander Shearer started the session by describing the use of such a database for

modeling an organism's responses to varied environments. Developing such flux-balance analysis (FBA) models is also a crucial test of quality that identifies gaps or errors in pathway annotations. Shearer described a gap-filling method to accelerate the building of FBA models by using a new tool, MetaFlux. The goal of this approach is to allow continuous process linking of an annotated genome to a model organism database, to a MetaFlux flux balance model and ultimately to new predictions. Eugenio Belda continued and this theme by describing the development of the MicroScope platform, a data structure that houses pathway annotations for large numbers of microorganisms and that incorporates tools to amalgamate curated results from diverse sources including a large body of community experts (4). Again, the importance was stressed of organizing the data structure to support a cyclical process in which accumulated data can be tested for consistency through modeling and the results fed back into improved annotation. Reannotation of *Bacillus subtilis* 168 as a test case resulted in assigning 6 new EC numbers and 17 UniProtKB/Swiss-Prot entry updates.

Most proteins are known only as predictions from whole-genome sequences. Assigning functions to newly described proteins based on sequence similarity with high confidence is thus critically important; Robert Finn and Marco Punta described related approaches to this problem. Finn and colleagues have developed a web-based application to build Hidden Markov Models from a user's data that is fast and incorporates a variety of displays and analysis features. The result of this exercise is a list of proteins ranked in order of their plausibility as members of a protein family. How should a quality threshold be set for membership? Setting a fixed threshold is appealing but results of Punta and colleagues indicate that no single threshold reliably excludes false positives from families. Some amount of manual biocuration is needed to yield optimal family groupings.

Constance Jeffrey discussed 'moonlighting' proteins, whose functions depart sharply from the ones predicted from their amino acid sequences. The best known example is perhaps the various lens crystallins, whose sequences are virtually identical to enzymes of intermediary metabolism. To find general ways to identify proteins with moonlighting potential, her group is systematically cataloguing physical and functional properties of known proteins, a bottom–up approach to structure–function annotation that complements the other approaches presented in the session.

Biocuration workflows and tools

Biocuration workflows and supporting tools vary considerably with the data type being curated. The presentations emphasized various aspects of the annotation process that are core values to the biocuration community: producing reusable tools, enforcing standards, improving annotation quality and consistency (peer-review or semi-automated approaches), and including text mining in the annotation pipeline. Greg Helt provided a preview of WebApollo, an open-source web-based genome annotation tool. Several features were demonstrated, including marking exon or intron edges to highlight support evidence and constructing an annotation model by dragging and dropping exons into the model being built. Attila Csordas described the Proteomics Identification database (PRIDE), a central archive of mass spectrometry and other proteomic data. This presentation included aspects of analysis and quality assurance workflows (stressing the need for these in the context of high-throughput data), public tools for data analysis and format conversions and integration of data with other resources such as UniProtKB (5). Julie Parks described CvManGO, a method for comparing computational- versus manual/literature-based GO annotation in the *Saccharomyces* Genome Database (SGD) that identifies discrepancies in GO annotations and can be used to help improve annotation quality (6). Marc Gillespie described the biocuration workflow for the Reactome pathway database. All entries are manually curated with content traceable to the primary literature. Entries are created though collaborations between Reactome annotators and domain experts, and undergo peer-review prior to public release. Details highlighted included the importance of a robust documentation framework for distributing public help documents, as well as close collaboration between the curators and reviewers. Ann Sarver gave a high-level description of the curation workflow at the Ingenuity Knowledge Base, a repository of protein interactions and functional annotations. The workflow leverages text mining and manual curation to generate 'expert findings' that are linked to publications and curated for accuracy.

Genomics, metagenomics, comparative genomics

Presentations in this session covered genome annotation tools, databases and reference datasets. Robert Riley presented the Joint Genome Institute's (JGI) web-based fungal genomics portal MycoCosm that integrates fungal genomics data and analytical tools and provides access to over 100 fungal genomes sequenced at JGI and elsewhere. Users may explore fungal genomes in the context of both comparative genomics and genome-centric analysis. MycoCosm promotes user community participation in data submission, annotation and analysis. Jennifer Harrow talked about the GENCODE consortium's aim to identify all gene features in the human genome, using a combination of computational and manual annotation approaches (7). She showed that the human transcriptome is far larger than originally thought, and the majority of this non-coding transcription has been classed as long non-coding RNA (lncRNA). The GENCODE 7 release contains 9640 lncRNA loci, including 3689 new loci. Of note, 3127

Database, Vol. 2012, Article ID bas036, doi:10.1093/database/bas036

of those new loci consist of two exon models indicating that they may be long non-coding loci. Aaron Mackey described ENIGMA, a tool that pools evidence across many gene predictors and EST/RNAseq data. Patrick Masson presented Viralzone, a web resource that contains comprehensive genomic information on viruses, including Baltimore classification, viral host, graphical displays of the virion structure and of its genome organization and descriptions of gene expression and replication. Dapeng Zhang presented a comparative genomic analysis that helped identify a new and widespread bacterial toxin system. The approach focused on identification of domains shared among components of bacterial toxin systems, as well as synteny. Raja Mazumder gave a presentation on the UniProt Representative Proteomes and Genomes effort. This provides a resource with a standardized set of proteomes and genomes ideal for use in genome annotation, metagenomic efforts and analyzing taxonomic nomenclature biases.

Protein structure, complexes, interactions

This session focused on the physical properties of proteins: their structures and their interactions, both with other proteins and with small molecules.

Two presentations described work to assess quality of models of protein structures. Juergen Haas presented recent developments in the Protein Model Portal (PMP) that support model validation and quality estimation, namely with the CAMEO tool (Continuous Automated Model EvaluatiOn). Marina Zhuravleva presented PDB's next generation validation reports that inform on structure–model quality and help identify potential problems. The reports will be made available to all interested users, particularly journal editors and peer reviewers.

Knowledge of protein–protein interactions is invaluable to help understand a protein's function and its regulation. Benjamin Shoemaker presented NCBI's Inferred Biomolecular Interaction Server (IBIS), which predicts interaction partners and locations of binding sites in proteins based on their evolutionary conservation in homologous structural complexes. IBIS provides binding site annotations for five different types of interaction partners (proteins, small molecules, nucleic acids, peptides and ions). It is estimated that about a third of the RefSeq sequences can be annotated with interaction partners using IBIS. Jyoti Khadake presented the IntAct editor, the curation tool used by the IntAct group and its collaborators. IntAct uses the Human Proteomics Organization's Proteomics Standards Initiative schema to store and exchange data. The tool is free and open-source.

Phoebe Roberts (Pfizer) presented targeted literature curation of therapeutic drug-induced toxic events. At Pfizer, scalable systems are developed to improve the quality of automatically extracted facts from literature. The focus is on entities and relationships of therapeutic interest, including targets, compounds, diseases and phenotypes, to understanding mechanistic underpinnings that lead to testable hypotheses. Extracted data are integrated with internal and external data sources for target evaluation, safety prediction and data analysis using computational approaches.

Jose Cruz-Toledo presented Aptamer Base. Aptamers are single-stranded nucleic acid or amino acid polymers that recognize and bind to targets with high affinity and selectivity. Aptamer Base is a database that provides detailed, structured information about the experimental conditions under which aptamers were selected and their binding affinity quantified. The database is being populated in a decentralized manner to keep up with new development in this area (8).

Integrating text mining in biocuration workflows

Several groups are working to help support biocuration by providing text mining tools to accelerate various aspects of the process. This session described recent developments in this area and was followed by a BioCreative workshop (Critical Assessment of Information Extraction in Biology); (Arighi *et al.*, submitted for publication).

Martin Krallinger described an experiment to elicit a systematic description of biocuration workflows from eight curation teams, as well as results from a survey of biocurator needs and experiences with text mining (9). This experiment was undertaken as a follow-up to a workshop held during the 2009 Biocuration Conference. The survey showed that, as of late 2009, half of the curators surveyed were using text mining in some part of the curation process. Most common uses of text mining are applications to improve prioritization of relevant documents for curation, identification of evidence (especially from full text) and linking of entities and relations to biological resources, e.g. EntrezGene or GO.

Two of the talks described tools that have been integrated into current biocuration workflows. Maximilian Haussler presented on annotating genomes with data from full text articles using a tool to extract genomic location information, including handling of pdf and other formats. The tool has been run over a large collection of full text articles from Elsevier and PubMedCentral. Using the extracted sequence information, a single curator was able to find 138 articles that confirmed *cis*-regulatory regions within 2.5 days. The tool is integrated into the University of California, Santa Cruz genome browser and is being used to annotate T-cell receptors. Kimberly Van Auken described an extension of the widely used Textpresso system to capture both GO Cellular Component and Molecular Function annotations. The approach combines statistical techniques to identify candidate papers containing relevant evidence, followed by use of Textpresso and Hidden Markov Models

(HMMs) to identify sentences and terms containing the desired molecular function relations for presentation to biocurators.

Two talks described experiments to validate text mining tools and adapt interfaces for specific curation needs. Fabio Rinaldi described the use of the ODIN system to validate extracted relations between drugs, genes and diseases from PharmGKB (10). The talk highlighted the need for repeated interactions and iteration with curators and the need for real data, in order to be able to adapt the system to curator needs. Daniel Jamieson described an experiment to recreate the HIV1–human protein interaction database using text mining techniques. The experiment demonstrated that it is possible to extract a large fraction of the relevant entities automatically, although event extraction was not as successful.

Ontologies and standards

The development of standards, be they of data exchange formats nomenclatures or reference sequences, has been a key focus of the new 'cooperative era' in the biomedical sciences. Accordingly, the talks given during the session on Ontologies and Standards either highlighted select go-to resources, or lent transparency to widely used procedures.

Marcus Chibucos presented the Evidence Code Ontology (ECO), including major changes to its structure: ECO now has two primary root classes, the evidence (including experimental assays, computational methods, author statements and inferences by biocurators) and the assertion method (i.e. manual or automated). He also highlighted how ECO can be used to document evidence in biological research. Jim Hu presented the Ontology for Microbial Phenotypes (OMP). The goal of this resource is to standardize the annotation of phenotypic information from bacteria and other microbes. Tobias Wittkop spoke about a web interface that allows researchers to perform term enrichment using over 200 ontologies, based upon the Annotator software created by the National Center for Biomedical Ontology (NCBO) that automatically annotates a gene or protein based on the corresponding Entrez Gene or UniProt textual description. Allen Davis followed with a talk describing the construction, implementation, maintenance and use of MEDIC, the disease vocabulary developed by the Comparative Toxicogenomics Database (CTD). MEDIC is a resource that integrates Online Mendelian Inheritance in Man (OMIM) terms, synonyms and identifiers with MeSH terms, synonyms, definitions, identifiers and hierarchical relationships (11). Kim Pruitt described the Consensus Coding Sequence (CCDS) project, is a collaboration between multiple centers with a goal of producing a set of high-quality protein coding region annotations for the human and mouse reference genome assemblies (12). The large number of available sequences in those species makes it very difficult for researchers to unambiguously describe the genes and proteins they are working on; therefore, efforts to integrate all the known coding sequences into a 'reference set' are essential. Alex Diehl described the development of the Neurological Disease Ontology (ND), an extension to the Ontology for General Medical Sciences (OGMS). John Anderson presented BioSample, a new NCBI resource that seeks to consolidate and unify source information for the data in NCBI's primary data archives.

Workshop 1: How to have a sustainable long-term plan for journals and databases?

This workshop consisted of a panel discussion on the interaction between databases and journals on the requirement for authors to provide meta-data for their submitted manuscripts in order to facilitate data integration in databases. This requirement is especially high for information provided as supplementary materials. For most data types, there are sufficient controlled vocabularies and ontologies available to define a standardized meta-data to describe published data. However, the establishment of a uniform specification will require significant effort by the journals and the scientific resource projects. The panel consisted of editors from four major journals; Thomas Lemberger (Chief Editor, Molecular Systems Biology, EMBO Journals), David Landsman (Editor in Chief, Database: The Journal of Biological Databases and Curation, Oxford University Press), Laurie Goodman (Editor in Chief, Giga Science), Michael Galperin (Executive Editor of the Nucleic Acids Research Database Issue, Oxford University Press), as well as Pascale Gaudet from the ISB; Michael Cherry and Francis Ouellette chaired the workshop. Gaudet represented the emerging standard BioDBCore to specifying meta-data for biological resources (http://biodbcore.org/) (13, 14). The policy stated by Galperin and Landsman requires the use of the BioDBCore for all databases described in papers published in DATABASE and Nucleic Acids Research Database issue.

GigaScience, a new online open access open data journal, has built a system that was designed expecting very large datasets. Similarly, the EMBO SourceData project aims to integrate data and structured metadata into papers. These initiatives will help ensure that raw data are preserved, reusable and discoverable. The panelists all seek a closer connection with the biocuration community to support biocuration and to facilitate the reuse of results from publications.

Workshop 2: Careers in biocuration

This workshop, chaired by Ilene Karsch Mizrachi and Monica Munoz-Torres, explored biocuration as a non-traditional career in the biological sciences. A majority of biocurators started their professional career as graduate and postgraduate research scientists in academic institutions, and later

Database, Vol. 2012, Article ID bas036, doi:10.1093/database/bas036

reoriented their careers to work in biocuration. The panelists were both from academia [Sarah Burge (Rfam); Beverly Underwood (NCBI)] and industry [Sam Ansari, (Philip Morris International; Jignesh Bhate, (Molecular Connections); Phoebe Roberts (Pfizer); Parthiban Srinivasan (Parthys Reverse Informatics)]. Sarah Burge discussed the findings of a survey of biocurators backgrounds, career paths and expectations (15); then panelists presented a brief overview of their career path and challenges associated with biocuration. Those presentations were followed by lively conversations about the priorities that must be set as a community to better train biocurators for the future. Participants and panelists concluded that it may be time for our community to actively conduct efforts to educate academic institutions on the importance of biocuration as a scientific career, and on the necessary special set of skills required of the curators.

Workshop 3: Quality information in support of annotations

As highlighted throughout the conference, common standards are of paramount importance to biological databases in order to make data exchangeable and reusable. Attribution of data provenance and evaluation of the quality of different data sources and methodologies is one area of biocuration where standardization efforts are greatly needed. The workshop on quality information to support annotations, chaired by Frederic Bastian and Marc Robinson-Rechavi [both from the Swiss Institute of Bioinformatics (SIB)] addressed this issue. The panelists [Marcus Chibucos (ECO), Michelle Giglio (ECO), Sylvain Poux (Swiss-Prot), Sandra Orchard (IntAct), Julio Collado-Vides (RegulonDB), Nives Skunca (OMA) and Suzanna Lewis (LBNL)] gave presentations highlighting how the resources they represent address annotations quality. It emerged that there are many varied systems to convey confidence information on annotations. Some groups have the users decide the quality of an annotation, whereas other groups try to provide some measure of the confidence. Possible uses and misuses of confidence information were debated. The GO uses ECOs that are sometimes incorrectly inferred to be indicative of quality. The workshop participants agreed that a different system needs to be developed. It was decided to create a working group to establish specifications for such a system, for instance, how to describe parameters used to assess the confidence of an annotation and defining a simple confidence score summarizing all the parameters. Work continues through a dedicated wiki: http://wiki.isb-sib.ch/biocuration/Quality_codes.

Workshop 4: Classification of diseases for curation of animal models

This workshop addressed an urgent topic for model organism databases and others seeking to improve the representations of the relationship of animal models to specific human diseases. Currently, for many of these groups, genetic diseases are represented by OMIM terminology but there are no clear solutions for the representation of common diseases or the relationships between them. The community needs a classification of disease not only useful for research purposes, but that also permits integration with currently accepted clinical terminologies and ontologies such as SNOMED-CT and ICD-10. A major need is a disease classification that will support structured access to animal models through their relationship to genetic diseases, the classic objective of model organism research.

It was agreed that in the future such a disease ontology would likely be radically different from those currently in use, and along the lines of the paradigm suggested by the recent report on precision medicine produced by the NAS (Committee on a Framework for Development a New Taxonomy of Disease, National Research C. Toward Precision Medicine: Building a Knowledge Network for Biomedical Research and a New Taxonomy of Disease: The National Academies Press 2012). Nevertheless, a pragmatic and functional resource is urgently required. Several panelists proposed various approaches aimed at addressing this issue, including MeSH [Olivier Bodenreider (NLM, Washington, DC, USA), MEDIC (Allan Davis, MDI-BioLabs, Mt. Desert, ME, USA)], SNOMED-CT, ICD-11 and UMLS were discussed. Also, the extent to which the existing Disease Ontologies [Lynn Schriml (Univ. MD School of Medicine, Institute for Genomic Sciences, Baltimore, MD, USA), Infectious Disease Ontology Linsay Cowell (University of Texas Southwestern Medical Center, Dallas, TX, USA)] or the Orphanet ontology of Mendelian diseases might provide a useful framework. Intense discussion among the 60 or so participants followed. As a result of this meeting, efforts are planned to coordinate the work of the groups represented, as well as other important contributors to this issue.

Workshop 5: NCBI and UniProt curation and tools

This session enabled the participants to understand some of the various activities at the NCBI and UniProt, and highlighted the close and mutually beneficial collaboration between them.

The UniProt presentations included an overview of the UniProt annotation workflow; the standards used in protein annotation; the curation of rules for propagation of annotation of uncharacterized proteins; the integration of genomics and proteomics information and the representation of complete proteomes. Sylvain Poux outlined the manual curation process, which consists of a review of the experimental data in the literature for each protein, the verification of the protein sequence and the annotation of the supporting evidence. Klemens Pichler presented the curation of rules in the UniRule automatic annotation

system and how they are used to enhance the annotation of a large number of poorly annotated protein sequences and invited participants to collaborate in the development of this project. Claire O'Donovan talked about the extensive cross-referencing in UniProtKB to more than 120 external databases that enables UniProt to provide core data for a particular protein with easy access provided to complementary data in external resources. The ongoing contact and active collaboration with external resource providers such as GenBank and the Model Organism Databases (MODs) ensures data quality and consistency. Maria Martin described the long-standing efforts of capturing complete proteomes, the recent release of Reference proteomes which are 'landmarks' in proteome space and explained how UniProt, Ensembl, ENA, GenBank and RefSeq work together to identify and maintain the complete proteome sets.

NCBI presented the flow of biological data from submission into the primary data archives, the steps taken during RefSeq curation, interactions with the community, annotation standards, application of pipelines and tools for validation and the interplay of human and machine curation. The steps taken during the indexing, and validation of data into the primary archives (GenBank) was presented by Ilene Karsch Mizrachi, including the automated validation steps, the different databases to which data flows, including BioProject, BioSample, GenBank and the Sequence Read Archive (SRA). RefSeq was the topic of the next three presentations, including eukaryotic genome and mRNA annotation and interactions with model organism databases by Melissa Landrum, prokaryotic annotation including work done on the model organism *Escherichia coli* K-12 and comparison of the annotation held in both NCBI and external databases, including UniProt, EcoGene and EcoCyc and protein family curation and naming comparison and incorporation of UniProt protein naming guidelines across RefSeq, UniProt, the Kyoto Encyclopedia of Genes and Genomes, and JCVI's TIGRFAMs by William Klimke. Rodney Brister discussed community annotation standards for viral genomes, engaging the community to obtain expert curation in order to seed annotation in protein clusters that can be used for further annotation propagation and resolving issues with respect to viral taxonomy through the International Committee on Taxonomy of Viruses. Finally, Tatiana Tatusova presented the results of NCBI's on-going annotation workshops that include experts in prokaryotic, viral, and fungal genomes, to set community-accepted annotation standards that can be used as validation checkpoints by the primary archives. A reannotation consortium composed of the NCBI, as well as major genome sequencing centers, The Broad Institute, JGI, JCVI, and IGS, was presented, that aims to generate consistent annotation for prokaryotic genomes, a critical need as NCBI expects to receive tens of thousands of clinical isolates for prokaryotic pathogens in the near future. This has led to the development of pan-genomic and additional resources for the analysis of multiple closely related genomes.

This session highlighted how value is added to biological data along the entire path, from automated validation tools all the way to highly intense manual curation efforts, engagement with the community in order to raise the annotation standards in a collaborative process and the on-going efforts to raise the bar higher every year as the amount of submitted data continues to grow.

Acknowledgements

The content of the conference was overseen by the Scientific Organizing committee, composed of Alex Bateman, Wellcome Trust Sanger Institute, UK; Judy Blake, Mouse Genome Informatics, USA; Mike Cherry, Saccharomyces Genome Database, USA; Pascale Gaudet, Swiss Institute of Bioinformatics, Switzerland; Michelle Giglio, University of Maryland, USA; Takashi Gojobori, National Institute of Genetics, Japan; Renate Kania, HITS gGmbH, Germany; Raja Mazumder, George Washington University, USA; Ilene Karsch Mizrachi, GenBank, USA; Monica C. Munoz-Torres, Georgetown University, USA; Claire O'Donovan, European Bioinformatics Institute, UK; Francis Ouellette, The Ontario Institute for Cancer Research; Kim D. Pruitt, RefSeq, USA; Bruno Sobral, Virginia Bioinformatics Institute, USA; Granger G. Sutton, JCVI, USA; Cathy Wu, Protein Information Resource, USA; Jasmine Young, PDB, USA.

The local arrangement committee provided excellent logistic support: Leslie Arminski, Kati Laiho, Peter McGarvey, Darren Natale, Thane Natarajan, Baris Suzek, and Sona Vasudevan from the Protein Information Resource, USA; and Raja Mazumder, and Monica C. Munoz-Torres from Georgetown University, USA.

We are also grateful to the Biocuration 2012 sponsors: Oxford University Press (Database: The Journal of Biological Databases and Curation and Bioinformatics), BioMed Central, The International Society for Biocuration, Georgetown University, Protein Information Resource, Department of Biochemistry and Molecular and Cellular Biology, The International ImMunoGeneTics information system, Protein Information Resource and Elsevier BV, Life Sciences.

Some of the talks and posters are available on slide share, under the keyword biocuration2012: http://goo.gl/Ps67j Iddo Friedberg blogged about the biocuration conference: http://bytesizebio.net/index.php/2012/04/06/biocuration-2012/.

The Biocuration 2012 Virtual Issue of the Database is available at http://www.oxfordjournals.org/our_journals/databa/biocuration_virtual_issue.html.

Funding

The publication fees of this article has been paid by the 2012 International Biocuration Conference committee and the International Society for Biocuration. This research was supported in part by the Intramural Research Program of the National Library of Medicine, NIH.

References

1. Stover,N.A., Punia,R.S., Bowen,M.S. *et al*. (2012) Tetrahymena Genome Database Wiki: a community-maintained model organism database. *Database*, **2012**, bas007.
2. Good,B.M., Clarke,E.L., Loguercio,S. *et al*. (2012) Building a biomedical semantic network in Wikipedia with Semantic Wiki Links. *Database*, **2012**, bar060.
3. Wang,Q., Arighi,C.N., King,B.L. *et al*. (2012) Community annotation and bioinformatics workforce development in concert–Little Skate Genome Annotation Workshops and Jamborees. *Database*, **2012**, bar064.
4. Vallenet,D., Engelen,S., Mornico,D. *et al*. (2009) MicroScope: a platform for microbial genome annotation and comparative genomics. *Database*, **2009**, bap021.
5. Csordas,A., Ovelleiro,D., Wang,R. *et al*. (2012) PRIDE: quality control in a proteomics data repository. *Database*, **2012**, bas004.
6. Park,J., Costanzo,M.C., Balakrishnan,R. *et al*. (2012) CvManGO, a method for leveraging computational predictions to improve literature-based Gene Ontology annotations. *Database*, **2012**, bas001.
7. Frankish,A., Mudge,J.M., Thomas,M. *et al*. (2012) The importance of identifying alternative splicing in vertebrate genome annotation. *Database*, **2012**, bas014.
8. Cruz-Toledo,J., McKeague,M., Zhang,X. *et al*. (2012) Aptamer Base: a collaborative knowledge base to describe aptamers and SELEX experiments. *Database*, **2012**, bas006.
9. Hirschman,L., Burns,G.A., Krallinger,M. *et al*. (2012) Text mining for the biocuration workflow. *Database*, **2012**, bas020.
10. Rinaldi,F., Clematide,S., Garten,Y. *et al*. (2012) Using ODIN for a PharmGKB revalidation experiment. *Database*, **2012**, bas021.
11. Davis,A.P., Wiegers,T.C., Rosenstein,M.C. *et al*. (2012) MEDIC: a practical disease vocabulary used at the Comparative Toxicogenomics Database. *Database*, **2012**, bar065.
12. Harte,R.A., Farrell,C.M., Loveland,J.E. *et al*. (2012) Tracking and coordinating an international curation effort for the CCDS Project. *Database*, **2012**, bas008.
13. Gaudet,P., Bairoch,A., Field,D. *et al*. (2011) Towards BioDBcore: a community-defined information specification for biological databases. *Nucleic Acids Res.*, **39**, D7–D10.
14. Gaudet,P., Bairoch,A., Field,D. *et al*. (2011) Towards BioDBcore: a community-defined information specification for biological databases. *Database*, **2011**, baq027.
15. Burge,S., Attwood,T.K., Bateman,A. *et al*. (2012) Biocurators and biocuration: surveying the 21st century challenges. *Database*, **2012**, bar059.