ISB International Society for Biocuration

14th Annual Biocuration Conference (virtual)

Abstracts from the invited speakers talks and poster session presentations

Contents

14th Annual biocuration conference (virtual) - overview

The 14th annual biocuration conference was virtual in 2021 due to the COVID-19 pandemic. We hosted four virtual sessions with talks and workshops, starting from April 13, 2021.

Session 1: 13 April 2021

Invited talks from Database Journal Virtual Issue

Speakers:

1. Patrick Ruch – *“A neural network based pipeline to classify publications for UniProtKB protein entries”*, HES-SO\HEG Geneva & SIB Swiss Institute of Bioinformatics
2. Constance Smith – *“Addition of RNA Seq Data and New Search Utilities to the Mouse Gene Expression Database (GXD)”*, The Jackson Laboratory, Bar Harbor, ME, USA
3. Randi Vita – “*A structured model for immune exposures”*, La Jolla Institute for Allergy and Immunology, San Diego, CA, USA
4. Federica Quaglia – “*APICURON: a database to credit and acknowledge the work of biocurators”*, CNR-IBIOM, University of Padova, Padova, Italy

Panel: Future of Biocuration

Panel Members:

1. Carol Bult – The Jackson Laboratory, Bar Harbor, ME, USA
2. James Malone – SciBite, Cambridge, UK
3. Kambiz Karimi – Myriad Women’s Health, South San Francisco, CA, USA
4. Sandra Orchard – EBI, Cambridge, UK

Session Chair: Rama Balakrishnan- Genentech, USA

Recording available here.

Session 2: 15 June 2021

Invited Speakers’ talks

Speakers:

1. Jiyu Chen – *“Automatic Consistency Assurance for Literature-based Gene Ontology Annotation*”, University of Melbourne, Australia.
2. Sushma Naithani – *“Biocuration training for undergraduate students: challenges and opportunities*”, Oregon State University, USA.

3. Livia Perfetto – *“A Resource for the Network Representation of Cell Perturbations Caused by SARS-CoV-2 Infection*”, Fondazione Human Technopole, Italy.
4. Sylvain Poux – *“Functional annotation of specific protein products in UniProtKB/Swiss-Prot*”, SIB Swiss Institute of Bioinformatics, Switzerland.

Panel: Career paths and projections in Biocuration

Panel Members:

1. Pankaj Jaiswal, Oregon State University, Oregon, USA
2. Tanya Berardini, Phoenix Bioinformatics, Fremont, CA, USA
3. Nicola Mulder, University of Cape Town, Cape Town, South Africa

Session chair: Peter Uetz, Virginia Commonwealth University, Virginia, USA

Recording available here.

Session 3: 17 August 2021

Workshop on Equity, Diversity and Inclusion (EDI)

Chair: *Sushma Naithani* – Associate Professor Senior Research & lead biocurator Plant Reactome, Oregon State University

Panel members:

1. Laurie Goodman – GigaScience Journal, Editor-in-Chief
2. Yasmin Alam-Faruque – Senior Biocurator at Healx
3. Varsha Khodiyar – Data Curation Manager at Springer Nature

Topics for discussion included also:

- the message captured in the movie Picture a Scientist (available on Netflix)
- better strategies to protect people from potential harassment and inequity in the workplace
- the role that professional societies, such as ISB, can play in promoting equity, diversity, and inclusion in our fields of study
- educating the community about manifestations of unconscious bias in our scientific fields

Invited Speakers’ talks

Speakers:

1. Nicholas Miliaras – *“Identifying New Chemical and Genetic Terms for Inclusion in the MeSH Vocabulary”*, National Library of Medicine Bethesda, Maryland, USA
2. Paula Duek – “*neXtProt function prediction community pages”*, Swiss Institute of Bioinformatics and University of Geneva, Switzerland

3. Jasmine Young – *“wwPDB Biocuration: On the Front Line of Structural Biology”*, RCSB Protein Data Bank, Rutgers, The State University of New Jersey, NJ, USA.
4. Anna Tanska – *“DrugMechDB: a database of drug mechanisms”*, Integrated Structural and Computational Biology Scripps Research, La Jolla, CA.

Session Chair: Federica Quaglia, National Research Council (CNR-IBIOM), University of Padova, Italy

Recording available here.

Session 4: 05 October 2021

Final conference session of the 2021 Virtual Conference series, featuring:

- 8:00 – 8:15 am PT: Annual General Meeting, Nicole Vasilevsky, Chair
- 8:15 – 9:00 am PT: Panel Discussion on Strategic Planning with former EC committee members:
    - Moni Munoz-Torres
    - Mike Cherry
    - Pascale Gaudet
    - Andrew Su
- 9:00 am – 10:00 am PT: Biocuration Award talks
    - Amos Bairoch
    - Anne Niknejad
- 10:00 am – 11:00 am – Poster session

Session Chair: Nicole Vasilevsky (ISB EC Chair), University of Colorado, USA

Recording available here.

This year’s conference sessions and poster session were sponsored by BioComputing UP lab (University of Padova, Italy). This project has received funding from the European Union’s Horizon 2020 research and innovation programme under grant agreement No 778247 and No 823886.

Useful links

ISB website

https://www.biocuration.org/

ISB YouTube

https://www.youtube.com/channel/UCNLZMHYSuWSIjoOinpAxo_Q

ISB Twitter

https://twitter.com/biocurator

ISB collection on F1000Research

https://f1000research.com/collections/biocuration

ISB2021 Virtual Conference – Posters on F1000Research

https://f1000research.com/collections/biocuration?years=2021&collectionId=98&selectedDomain=slides

ISB2021 Virtual Conference – Slides from Invited Speakers' talks on F1000Research

https://f1000research.com/collections/biocuration?years=2021&collectionId=98&selectedDomain=posters

Abstracts

Abstracts of the 14th annual biocuration conference of 2021, including the *Invited Speakers' talks* abstracts (2nd and 3rd sessions) along with the abstracts from the *Poster Session (October 5th)*.

Biocuration training for undergraduate students: challenges and opportunities

Sushma Naithani[1]

[1] Department of Botany and Plant Pathology, Oregon State University, Corvallis, OR, 97333, USA

ISB2021 session: session 2, June 15th

Category: Curation and Data Visualization Tools

Plant genomes harbor unique complexities such as polyploidy and extensive duplication of genomic regions and genes. However, the annotation of plant genes still relies on gene-orthology based projection from model species providing identical GO annotations to all gene family members, with a large number of remaining genes sifted into unknown category. We implement biocuration to improve the functional annotation of gene family members based on published scientific literature, analysis of expression patterns of homologous genes, and evidence from gene-gene interaction data sets. In these efforts, we engage undergraduate students in small research projects designed to develop their genomic data literacy, and to support biocuration efforts of plant genes and pathways by leveraging affiliated bioinformatics data science courses, thesis projects, research programs, and experiential summer learning internships for undergraduates—all available at our institution. Notably, two species-specific, metabolic pathway databases (FragariaCyc and VitisCyc) were curated solely with help from undergraduate students, a set of Python scripts for capturing orthology projection statistics from the Plant Reactome knowledgebase were written by an undergraduate REEU summer intern, and the functional annotation of 144 rice genes belonging to the S-domain related receptor-like kinase (SDRLK) gene family was improved as part of an undergraduate honors thesis. Additionally, we hosted jamborees for genes and pathway biocuration to encourage educators and plant researchers and graduate students to become a community biocurator. The participants acquire Big Data literacy and additional analytical skills useful for conducting and publishing their own research. We will discuss the challenges, opportunities, strategies for successful outcomes of these projects. We acknowledge Oregon State University funds to SN, NSF award IOS #1127112 to Gramene project, and NIFA award #2019-67032-29072 on Ag-REEU: Undergraduate Learning Experiences in working with Big Data in Agriculture.

A Resource for the Network Representation of Cell Perturbations Caused by SARS-CoV-2 Infection

Livia Perfetto[1]

[1] Department of Biology, Fondazione Human Technopole, Milan, Italy

ISB2021 session: session 2, June 15th

Category: Curation and Data Visualization Tools; Human and Environmental Health

The coronavirus disease 2019 (COVID-19) pandemic has caused millions of casualties worldwide and the lack of effective treatments is a major health concern. The development of targeted drugs is held back due to a limited understanding of the molecular mechanisms underlying the perturbation of cell physiology observed after viral infection. Recently, several approaches, aimed at identifying cellular proteins that may contribute to COVID-19 pathology, have been reported. Albeit valuable, this information offers limited mechanistic insight as these efforts have produced long lists of cellular proteins, the majority of which are not annotated to any cellular pathway. We have embarked in a project aimed at bridging this mechanistic gap by developing a new bioinformatic approach to estimate the functional distance between a subset of proteins and a list of pathways. A comprehensive literature search allowed us to annotate, in the SIGNOR 2.0 resource, causal information underlying the main molecular mechanisms through which severe acute respiratory syndrome coronavirus 2 (SARS-CoV-2) and related coronaviruses affect the host–cell physiology. Next, we developed a new strategy that enabled us to link SARS-CoV-2 interacting proteins to cellular phenotypes via paths of causal relationships. Remarkably, the extensive information about inhibitors of signaling proteins annotated in SIGNOR 2.0 makes it possible to formulate new potential therapeutic strategies. The proposed approach, which is generally applicable, generated a literature-based causal network that can be used as a framework to formulate informed mechanistic hypotheses on COVID-19 etiology and pathology.

Functional annotation of specific protein products in UniProtKB/Swiss-Prot

Sylvain Poux[1]

[1] SwissProt group, SIB Swiss Institute of Bioinformatics, Geneva, Switzerland

ISB2021 session: session 2, June 15th

Category: Curation and Data Visualization Tools; Functional Annotation

The UniProt Knowledgebase (UniProtKB, https://www.uniprot.org) is a comprehensive, high quality and freely accessible resource of protein sequences and functional information. UniProtKB is composed of an expert curated section, UniProtKB/Swiss-Prot, and its automatically annotated complement UniProtKB/TrEMBL. A UniProtKB/Swiss-Prot record provides expert curated protein sequences for all transcripts expressed by one gene. Sometimes the functions of such isoform sequences may be as different from each other as the functions of the products of distinct genes.

Here we describe recent developments that allow UniProt curators to capture the distinct functions and features of individual protein isoforms and mature chains in UniProtKB/Swiss-Prot. These “isoform/chain-specific annotations” will provide better support for functional interpretation of transcriptomics and proteomics data, particularly from single cells, and for machine learning approaches that attempt to relate sequence features to any aspect of protein biology – be it protein function, interactions, distribution, or roles in disease. This work forms part of a wider strategy that aims to improve the machine readability of curated knowledge in UniProtKB/Swiss-Prot more generally, which also encompasses the standardization of small molecule interactions using Rhea and the functional and clinical significance of human variation data (both of which will also be presented at this meeting).

UniProtKB/Swiss-Prot currently provides functional annotations for specific isoforms and/or chains for more than 5,000 proteins, including more than 1,000 of human origin.

Automatic Consistency Assurance for Literature-based Gene Ontology Annotation

Jiyu Chen[1], Nicholas Geard[1], Justin Zobel[1], Karin Verspoor[2]
[1] University of Melbourne, Australia
[2] RMIT University, Australia

ISB2021 session: session 2, June 15th

Category: Curation and Data Visualization Tools, Data Standards and Ontologies, Functional Annotation

Literature-based gene ontology (GO) annotation is a process where expert curators use uniform expressions to describe gene functions reported in research papers, creating computable representations of information about biological systems. Manual assurance of consistency between GO annotations and the associated evidence texts identified by expert curators is reliable but time-consuming and is infeasible in the context of rapidly growing biological literature. A key challenge is maintaining consistency of existing GO annotations as new studies are published and the GO vocabulary is updated.

In this work, we introduce a formalisation of biological database annotation inconsistencies, identifying four distinct types of inconsistency. We propose a novel and efficient method using state-of-the-art text mining models to automatically distinguish between consistent GO annotation and the different types of inconsistent GO annotation. We evaluate this method using a synthetic dataset generated by directed manipulation of instances in an existing corpus, BC4GO.

Two models built using our method for distinct annotation consistency identification tasks achieved high precision and were robust to updates in the GO vocabulary. We provide detailed error analysis for demonstrating that the method achieves high precision on more confident predictions. Our approach demonstrates clear value for human-in-the-loop curation scenarios.

1) database curators, gene ontology annotation curators, bioinformatics, BioNLPer
2) learning objective
a. introduce a formalisation of biological database annotation inconsistencies, identifying four distinct types of inconsistency, address the key challenge of maintaining the consistency of existing GO annotations as new studies are published and the GO vocabulary is updated.

b. propose a novel and efficient solution to the challenge using state-of-the-art text mining models to automatically distinguish between consistent GO annotation and the different types of inconsistent GO annotation

wwPDB Biocuration: On the Front Line of Structural Biology

Jasmine Young[1]

[1] Research Collaboratory for Structural Bioinformatics Protein Data Bank (RCSB PDB), Institute for Quantitative Biomedicine, Rutgers, The State University of New Jersey, Piscataway, NJ, USA

ISB2021 session: session 3, August 17th

Category: Data Standards and Ontologies

The Protein Data Bank (PDB) provides open access to >175,000 biological macromolecules that facilitate understanding of the structure and its function in biology. wwPDB collaboration and expert biocuration is a key to provide the best-curated biological data resource. wwPDB biocurators are at the very front line of structural biology who bridge between the scientists who submit PDB structures and the public who use these data. Our ultimate goal is to support research, training, and education worldwide and facilitate public awareness and understanding of these scientific discoveries.

neXtProt function prediction community pages

Paula Duek[1,2], Camille Mary[2], Monique Zahn[1], Amos Bairoch[1,2], Lydie Lane[1,2], and the neXtProt team

[1] CALIPHO group, SIB Swiss Institute of Bioinformatics

[2] Department of microbiology and molecular medicine, Faculty of medicine, University of Geneva, Geneva, Switzerland

ISB2021 session: session 3, August 17th

Category: Functional Annotation

Around 90% of the ~20'000 predicted human protein coding genes have functional annotations in neXtProt. The HUPO Human Proteome Project (HPP) has recently launched a project aiming to characterize the remaining 10% of human proteins that have no annotated function. To support this project and the scientific community in its efforts to complete the human functional proteome, neXtProt proposes community pages where functional predictions on human proteins, standardized using GO and ECO terms, can be shared. The submitter(s) can remain anonymous or have their ORCID(s) linked to the prediction to give credit and promote exchange with interested researchers. neXtProt started to populate the function prediction pages with functional hypotheses obtained in the frame of the Fonctionathon course for undergraduates given at the University of Geneva.
For an example, see https://www.nextprot.org/entry/NX_Q6P2H8/function-predictions. The students extracted data from literature and databases and used bioinformatic tools to formulate their functional predictions. We are inviting researchers and biocurators who daily do manual exploration to propose their functional predictions. We welcome suggestions for improvements on the new neXtProt community pages.

DrugMechDB: a database of drug mechanisms

Anna Tanska, Mike Mayers, Denise Silva, Umasri Sankarlal, Lakshmanan Jagannathan, Patrick Rewers, Andrew I. Su
Scripps Research, Department of Integrative Structural and Computational Biology, La Jolla, CA

ISB2021 session: session 3, August 17th

Category: Curation and Data Visualization Tools

There are many databases of drug indications that annotate chemical compounds and the diseases that they treat. However, annotation of the specific mechanism of action is generally limited to broad mechanistic categories and/or protein targets. Here, we describe DrugMechDB, a database of drug mechanisms that describes drug-disease indications as paths through a biomedical knowledge graph. Intermediate nodes between the drugs and the diseases they treat will include nodes for pathways, phenotypes, genes and proteins, Gene Ontology terms, and metabolites. DrugMechDB is available at https://sulab.github.io/DrugMechDB/.

Identifying New Chemical and Genetic Terms for Inclusion in the MeSH Vocabulary

Nicholas Miliaras[1], Joseph Denicola[2], Justine Fitzgerald[3], Stella Koppel[1], Shari Mohary[1], Oleg Rodionov[1], Olga Printseva[1], James Pash[2]

[1] Index Section, National Library of Medicine, Building 38, 8600 Rockville Pike, Bethesda, Maryland 20894, USA

[2] Medlars Management Section, National Library of Medicine, Building 38, 8600 Rockville Pike, Bethesda, Maryland 20894, USA

[3] Medical Science & Computing LLC, 11300 Rockville Pike, Suite 1100 Rockville, MD 20852 USA

ISB2021 session: session 3, August 17th

Category: Data Standards and Ontologies

Indexing is the process of assigning terms from a controlled vocabulary to facilitate search and retrieval of relevant text. Medical Subject Headings (MeSH), a controlled vocabulary used by the National Library of Medicine, is critical for the indexing of PubMed literature, yet indexers regularly encounter new terms for drugs, chemicals, or genes that have not been entered in MeSH. In order to make such terms available for indexing and facilitate searches for these topics, we have established a process where indexers identify or “flag” new terms for review and possible inclusion in MeSH. Criteria for inclusion in MeSH are: (1) Is the term not found as a synonym in MeSH? (2) For chemicals and drugs, does the term occur in more than one article? (3) For drugs, if the term occurs only once, is the article based on a clinical trial? (4) For proteins and microRNAs, does the term represent one of the organisms established by MeSH as model research organisms, or a clinically-relevant pathogen? For terms that satisfy any of these criteria, we create a new MeSH record, or add the term to an existing record where it becomes a synonym, and the article is indexed with the new term. However, for chemical or drug terms that occur only once in the literature, we place these terms on hold for future searches. We present the flag workflow, data from the process, and describe a method we use to search for new publication of terms placed on hold.

Automatic literature mining tool to extract glycosylation information from literature

Rahi Navelkar[1]

[1] The Department of Biochemistry and Molecular Biology, George Washington University Medical Center, Washington, DC 20037

ISB2021 session: session 4, October 5th

Category: Curation and Data Visualization Tools, Automatic literature mining tool

Glycosylation is one of the most common and complex post-translation phenomena which impacts several key biological processes making it vital to study it with regard to human health and disease. Recent years have seen a great influx of data in glycobiology research with many papers reporting glycosylation sites, proteins, differential glycosylation in reference to disease, etc. However, transferring such data into existing bioinformatics databases for public use requires manual curation by experts, which is often time-consuming and expensive. Accelerating this process is key in facilitating biomedical research by providing the latest findings in a standardized way, ready for research use.

To facilitate bio-curation, we developed a literature mining tool that will detect glycosylated proteins, sites and glycan descriptions automatically in abstracts of publications from the MEDLINE resource. Extracted protein names are normalized to their NCBI gene ID and UniProtKB accession and the data is processed through manual and automatic quality control (QC) checks. The QC process assesses the validity of the protein accessions as well as the reported sites against fasta sequence from the UniProtKB database. The data that passes all QC checks is integrated in the GlyGen (https://www.glygen.org/) database and is publicly available.

Future plans include fine-tuning the tool to detect papers describing large-scale experiments (with mentions over 20 glyco-proteins). Papers identified will be manually curated by experts that contribute to GlyConnect (glyconnect.expasy.org) a partner resource of GlyGen. Finally, we plan to dockerize the system and run it periodically on all “MEDLINE” abstracts to extract relevant information from the latest literature. The generated data will be fed to GlyGen’s data integration pipeline where users can perform complex queries and download the data. The extracted data can be accessed through GlyGen data (https://data.glygen.org/GLY_000481) or through individual protein pages (e.g https://www.glygen.org/protein/P01106-1#Glycosylation (under Text Mining tab)).

Current Challenges in Optimizing the Capture of all Publications Pertaining to a Protein Family and its Members

Colbie Reed[1], Rémi Denise[1], Geoffrey Hutinet[1], and Valérie de Crécy-Lagard[1]

[1] Microbiology and Cell Science Department, The University of Florida, Gainesville, Florida 32603

ISB2021 session: session 4, October 5th

Category: Comparative Genomics; Data Standards and Ontologies; Functional Annotation; real-world use of available resources; UX standards, performance and impact

Although conceptually simple, one of the major challenges remaining in the post-genomic era is the capture of all published experimental data available for members of a protein family. To date, two main methods are used to search the scientific corpus for works pertaining to a given sequence and its homologs: 1) text-based search using common family/homolog identifiers; 2) sequence-based search using a member sequence to query a single or several databases to identify existing links between published data and matching homologs. Here, select tools commonly used to capture the literature on a protein family are reviewed and current challenges discussed. Using DUF34 as an example protein family, a defined list of keywords and a subset of member sequences were determined for use in systematic text- and sequence-based queries, respectively. Tools and their performances were then evaluated and compared, revealing major differences in the number and quality of the identified publications. In addition to highlighting tool vulnerabilities as to better inform the common practices of researchers and their use of such resources, these results provided insights important for addressing key challenges of sequence-to-publication crosslinking, biological entity identification, and data publishing standards.

Measuring Success and Scientific Impact -- Human Disease Ontology

Lynn M. Schriml[1] and J. Allen Baron[1]

[1] Institute for Genome Sciences, University of Maryland School of Medicine

ISB2021 session: session 4, October 5th

Category: Data Standards and Ontologies

With the Human Disease Ontology (DO) serving as the disease nomenclature resource across more than 280 biomedical resources, including more than 50 biomedical ontologies, the bulk of the Model Organism Databases (over 195,000 annotations of DO diseases to genes, alleles, or genomic models among the Alliance of Genome Resources), and the NIH Common Fund Data Ecosystem, identifying how, by whom, and for what purposes DO is utilized has become a significant challenge. Tracking these metrics in order to discern the breadth of our scientific community and scientific impact is key to identify areas of growth, substantiating utility and informing future development.

The Human Disease Ontology, established in 2003, is a NHGRI-funded Genomic Resource that provides the biomedical community with a comprehensive, expertly curated, computationally tractable disease knowledgebase. The DO semantically classifies the breadth of human diseases (10,855 diseases) and integrates disease and clinical vocabularies through extensive cross mapping of DO terms by both automated methods, via biannual updates of the UMLS (including to MeSH, ICD-9, ICD-10, NCI's thesaurus, SNOMED CT), and through manual curation of OMIM, GARD, Orphanet and ICD-O. The DO provides a stable framework for advanced analysis of disease through curated and ROBOT-automated import of non-disease ontologies to define logical axioms describing anatomical, genetic, clinical, and environmental factors associated with disease.

Here we describe approaches to identify, measure, and report on DO's impact and user community along with recommendations for best practices and utilization of these approaches by other resources. These approaches include: 1) comparison of various resources for tracking citations (PubMed, Europe PMC, Scopus, Google Scholar); 2) how to capture publications that use but do not cite a resource (a common problem for established biomedical resources); 3) methods for determining how a resource is used; 4) methods for documenting usage; and 5) discussion of manual and automated approaches (e.g. APIs, R packages). Our goal is to aid other resource developers in identifying their impact and user communities and to encourage open dialog and idea generation for future improvement towards measuring success and scientific impact of our biomedical resources.

DisProt in 2021: expanding manually curated annotation of IDPs

Edoardo Salladini[1], Federica Quaglia[1,2], András Hatos[1], Silvio C.E. Tosatto[1] and Damiano Piovesan[1]

[1] Department of Biomedical Sciences, University of Padova, Padova, Italy

[2] Institute of Biomembranes, Bioenergetics and Molecular Biotechnologies, National Research Council (CNR-IBIOM), Bari, Italy

ISB2021 session: session 4, October 5th

Category: Curation and Data Visualization Tools; Data Standards and Ontologies; Functional Annotation

In 2020 DisProt (https://disprot.org/), the manually curated database of intrinsically disordered proteins (IDPs), transitioned to a more frequent – six-month based – release schedule. The last releases feature newly added proteins as well as revised and updated published entries. Several aspects of IDPs are now extensively covered and explored by a small team of expert DisProt reviewers, in an effort to continuously improve curation accuracy and provide comprehensive and up-to-date annotations. Entries describing IDRs in SARS-CoV-2 have been added as a response to the global pandemic, e.g. the SARS-CoV-2 Spike glycoprotein (DP02772) and ORF3a protein (DP03003). To aid comparative studies, we also extended our analysis to the Spike glycoproteins of two closely related beta coronaviruses that caused outbreaks of severe respiratory illnesses in humans: MERS-CoV and SARS-CoV. One of our main goals at DisProt is to continuously review and update entries of intrinsically disordered proteins in light of new results being published, to ensure that we provide the most comprehensive annotations. To this end we revised and updated several well-known IDPs including HMGA1 (DP00040), HSP12 (DP00705), Neuromodulin (DP00955), p27 (DP00018). Moreover, we introduced the first two thematic DisProt dataset "unicellular toxins and antitoxins"and “extracellular matrix proteins”, easily accessible from the DisProt home page. All the entries included in these datasets have a tag attached to help users easily identify them, as well as having the option to download them in a single file. Several other biological areas where IDPs play a crucial role will be covered and described with new datasets, to be published in the next releases of the database. Finally we are now adopting two stable ontologies - Gene Ontology (GO) and the Evidence and Conclusion Ontology (ECO) - to improve interoperability and data standardization of our manually curated annotations.

Reclassification and Axiomatization of Chromosomal Anomaly Diseases in the Mondo Ontology

Sabrina Toro[1], Nico Matentzoglu[2], Melissa Haendel[1], Chris Mungall[3], Nicole Vasilevsky[1]

[1] University of Colorado Denver, Anschutz Medical Campus, Aurora, CO USA

[2] semanticly, Athens, Greece

[3] Lawrence Berkeley National Laboratory, Berkeley, CA, USA

ISB2021 session: session 4, October 5th

Category: Data Standards and Ontologies

The Mondo Disease Ontology (Mondo) integrates multiple disease terminologies into a coherent logic-based ontology that provides precise semantic mappings between terms. It harmonizes several source terminologies and ontologies covering various aspects of disease across species and includes cross-references to the other databases using strict semantic allowing for computational integration of disease associated data (such as phenotypes associated with a disease). Mondo uses Dead Simple Ontology Design Patterns to consistently and automatically apply equivalence and subclass axioms to categories of diseases, allowing automatic classification and inferencing.
The INCLUDE Data Coordinating Center (https://includedcc.org/) provides access to data, analysis tools, and resources for the Down syndrome community, and standardizes and harmonizes Down Syndrome patient data using Mondo. This group, and others like it, requires proper classification of chromosomal anomaly terms, for instance, proper classification of Down Syndrome and children terms.
Here we report the creation of patterns to support the reclassification of the chromosomal anomaly branch of Mondo.

'Chromosomal anomaly' was reclassified based on the following concepts: chromosome number anomaly (aneuploidy and polyploidy), chromosome structure anomaly (e.g. deletion or duplication of part of the chromosome), and anomaly based on chromosome type (e.g. autosomal anomaly and gonosomal anomaly).

Patterns for each concept of 'chromosomal anomaly' were created using axioms representing the type of chromosomal anomaly and the chromosome/chromosome region at the root of the disease.
These axioms are, respectively, 'has modifier' some chromosome_variation (SO:0000240), and 'disease arises from structure' some (chromosome (GO:0005694) or 'chromosomal region' (GO:0098687))
The 'chromosome' and 'chromosome region' are terms of the newly created ontology Monochrom which was created to convert chromosomes and chromosome bands data from UCSC Genome Browser into an OWL classification and therefore allow for these terms to be used in axioms.

This ongoing work will ensure consistency and automatic classification of chromosomal anomaly terms in Mondo, which will support the work of Mondo users such as INCLUDE Data Coordinating Center.
Information about Mondo can be found here: https://github.com/monarch-initiative/mondo

The Xenopus Phenotype Ontology on Xenbase

Malcolm Fisher[1], Erik Segerdell[1], Joshua D. Fortriede[1], Christina James-Zorn[1], Mardi J Nenni[1], Troy Pells[2], Virgilio Ponferrada[1], Nivitha Sundararaj[1], Kamran Karimi[2], Praneet Chaturvedi[1], Vaneet Lotay[2], Stanley Chu[2], Joe Wang[2], Eugene Kim[2], Sergei Agalakov[2], Brad Arshinoff[2], Peter D. Vize[2], and Aaron M. Zorn[1]

[1] Division of Developmental Biology, Cincinnati Children's Hospital, Cincinnati, OH, United States

[2] Departments of Biological Science & Computer Science, University of Calgary, Calgary, Alberta, Canada

ISB2021 session: session 4, October 5th

Category: Curation and Data Visualization Tools; Data Standards and Ontologies

Xenbase, the Xenopus database, continues to develop tools and support to accelerate basic research and human disease modeling. A major goal of Xenbase is to make the large amount of data generated by Xenopus research more accessible to the research community. Our latest tool to achieve this goal is the ‘Xenopus Phenotype Ontology’ (XPO), an ontology designed around principles laid out for cross species interoperability. The XPO has terms covering a broad spectrum of phenotypes including gross adult morphology, early development, behavior and subcellular components. Xenbase now routinely uses the XPO as part of our phenotype curation process, alongside curation of expression phenotypes as statements composed of ontology terms and Xenbase genes using a tightly constrained syntax. XPO based phenotypes and expression phenotypes, along with putative modelled diseases in the form of Disease Ontology (DO) terms, are linked to summary descriptions of the experiments that generated them, built using a set of controlled vocabularies to improve searchability. We also make use of previously generated lists of differentially expressed genes that are a product of our GEO sequencing data pipeline, where we take a conservative approach to produce up and down regulated gene expression phenotype statements. All of this new phenotype data is integrated into our standard article page views, as annotations for specific images, as well as on new phenotype pages. The new access points include a phenotype search and phenotype tabs on gene pages which collate and categorize all phenotype data related to that specific gene. This growing body of phenotype data, along with increased scope for cross species comparisons, should allow us to make it easier for users to gain insights into human disease from existing Xenopus data, as well as to suggest possible new disease models. Xenbase is funded by a Biotechnology Resource grant from the NICHD.

Expanding and Enriching the Representation of Alleles in the Alliance of Genome Resources

Susan M Bello[1], Yvonne M Bradford[2], Cynthia L Smith[1], and the Alliance of Genome Resources Disease and Phenotype working group

[1] The Jackson Laboratory, Bar Harbor ME

[2] University of Oregon, Eugene OR

ISB2021 session: session 4, October 5th

Category: Data Standards and Ontologies

The Alliance of Genome Resources (Alliance) has been working to harmonize the representation of alleles across the member model organism databases (MODs). In the first phase, the focus was on harmonizing and incorporating the minimal set of essential features used in single gene allele annotations including; symbol, CURIE, allele_of relation to a gene, description of the molecular mutation. In the second phase, we expanded the allele representation beyond just single gene alleles to also include transgenic alleles. Transgenic alleles differ from single gene alleles in that they do not require an allele_of relation to a gene and may have a relation to a construct. Constructs were added to the Alliance to capture the features of the transgenic vectors used to create transmissible transgenic alleles. Constructs include fields for documenting the promoters and enhancers used to drive expression and the genes expressed. CURIEs are included for the genes used to supply the promoters, enhancers, and expressed gene sequences whenever available. Inclusion of these CURIEs allows for display of transgenic alleles expressing a gene on relevant Alliance gene pages in a transgenic allele table regardless of the species carrying the transgenic allele. Constructs may be related to zero to many transgenic alleles and transgenic alleles may have relationships to zero to many constructs. In the next phase we are working to further expand the information associated with alleles to cover all types of information annotated by any of the participating MODs. Where possible, we are harmonizing existing MOD specific controlled vocabularies and working to get these terms incorporated in public ontologies. For example, the generation method used to create or identify the alleles is annotated using in-house controlled vocabularies or ontologies by many of the MODs. Thus in MGI and FlyBase there are a terms for all endonuclease techniques - “endonuclease-mediated” in MGI and “site specific cleavage” (FBcv:0003007) in FlyBase - while in FlyBase and ZFIN there are individual terms for the different techniques using endonucleases (e.g. CRISPR/cas, TALENs). From the various in-house vocabularies a consensus list of terms was created and these terms were submitted to the Ontology for Biomedical Investigations (OBI). Work to harmonize additional vocabularies, bring in alleles affecting multiple genes, and incorporate the full range of information about alleles into the Alliance is ongoing.

Biomappings: Community Curation of Mappings between Biomedical Entities

Charles Tapley Hoyt[1], Amelia L. Hoyt[2], Benjamin M. Gyori[1]

[1] Laboratory of Systems Pharmacology, Harvard Medical School, Boston, MA 02115, USA

[2] Independent researcher

ISB2021 session: session 4, October 5th

Category: Data Standards and Ontologies

Many related biomedical resources propose their own identifiers for genes, proteins, chemicals, biological processes, and other entities of biological interest. The integration of data and knowledge fundamentally relies on mappings between equivalent entities across resources. While some maintain and distribute mappings to external namespaces (e.g. HGNC provides mappings to Entrez Gene identifiers), there exist systematic gaps in the availability of mappings between widely used resources. Automated approaches including lexical and structural alignment have been used to find missing mappings, but most do not store important metadata like mapping confidence nor important curation artifacts including proposed mappings curated to be (nontrivially) incorrect, nor provide interfaces for curating (reviewing, and confirming or rejecting) predicted mappings.

We introduce Biomappings, an open repository for making expert-curated mappings (currently 5,700), and predicted mappings (currently 28,000) available with their associated metadata in an intuitive tab-delimited format. It is licensed under the permissive CC0 license to encourage community contributions and restriction-free integration back into primary resources. Biomappings provides a web-based interface for curating predictions and adding manually constructed mappings, a Python package for interacting with the mappings, and several workflow examples for generating new mappings. We applied the Biomappings curation workflow to missing mappings between the Medical Subject Headings and several other ontologies including the Disease Ontology, ChEBI, and the Gene Ontology. We also used Biomappings to curate an exhaustive set of proposed mappings (constructed automatically based on lexical overlap) between entries representing cancer cell lines across three previously unmapped resources: the Cancer Cell Line Encyclopedia, the Experimental Factor Ontology, and Cellosaurus.

All data and code are available at https://github.com/biopragmatics/biomappings

Characterizing novel protein (domain) families from proteome “dark matter”

Aron Marchler-Bauer, Myra Derbyshire, Noreen Gonzales, Marc Gwadz, Chanjuan Zheng
National Center for Biotechnology Information, National Library of Medicine, Bethesda, MD 20894, USA

ISB2021 session: session 4, October 5th

Category: Curation and Data Visualization Tools, Functional Annotation

We present two overlapping approaches for identifying protein and protein domain families from sets of representative sequences that currently lack annotation and classification: (1) building models for novel families around experimentally determined 3D structures, and (2) building models around sequences with functional information available from the published literature. These strategies yield rich sets of family model candidates. We report on the results of a study relating curation effort to outcome as measured by increase in sequence coverage. Conserved Domain Database (CDD) curators selected candidate models, refined pre-computed alignments, and abstracted functional information. A set of about 600 novel models provided coverage for more than 2% of sequences from a representative set of 2.3 million. However, the fraction of sequences uniquely annotated by the new models is smaller by an order of magnitude. We observe that novel domain models often complement older models in characterizing more complete domain architectures, and that novel domains often appear to be members of previously characterized families, underscored by structural similarity where experimentally determined 3D structure is available.

UniProtKB and Neurodegenerative Diseases: Towards a deeper understanding of underlying mechanisms

Yvonne C. Lussi[1], Elena Speretta[1], Kate Warner[1], Michele Magrane[1], Sandra Orchard[1] and UniProt Consortium[1,2,3]

[1] European Molecular Biology Laboratory, European Bioinformatics Institute (EMBL-EBI), UK

[2] SIB Swiss Institute of Bioinformatics, Centre Medical Universitaire, Switzerland.

[3] Protein Information Resource, Georgetown University, USA

ISB2021 session: session 4, October 5th

Category: Functional Annotation, Human and Environmental Health

Neurodegenerative diseases are a heterogeneous group of disorders that are characterized by the progressive degeneration of the structure and function of the central or peripheral nervous systems. Although treatments may help relieve some of the physical or mental symptoms associated with neurodegenerative diseases, there is currently no way to slow disease progression and no known cures. In order to find treatments for these incurable and debilitating disorders, it is crucial to understand the underlying mechanisms that lead to disease development and identify the genetic variants that are associated with the disease. In this context, UniProtKB aims to link genetic and medical information to protein sequences and associated biological knowledge to improve our understanding of disease development.

We focus on a thorough review of available information on sequence variants and associated neurodegenerative disease information linking to specialized databases, as well as collecting literature information on the normal function of proteins associated with the disease. The information on variants together with variant functional description, protein molecular function, structural data and protein-protein interactions should help researchers in the field of neurodegeneration, clinicians and biomedical researchers to gain a global view on the relation between variant and disease and help in elucidating disease mechanism.

To further facilitate retrieval of disease-specific information, we will follow previous efforts at UniProKB which focused on the annotation of information on proteins and protein variants involved in Alzheimer disease. These efforts included the development of the Disease Portal, which is a disease-centric entry point into UniProtKB, supporting the navigation through information available on Alzheimer disease. The aim will be to expand the information exhibited in the Disease Portal to additional neurodegenerative diseases.

Curating clinical phenotypes in PHACE Syndrome to reveal disease mechanisms

Nicole Vasilevsky[1]

[1] University of Colorado Denver, Anschutz Medical Campus, Aurora, CO USA

ISB2021 session: session 4, October 5th

Category: Human and Environmental Health

PHACE syndrome is a rare disorder of unknown etiology characterized by the association of Posterior fossa brain malformations, large facial Haemangiomas, cerebral Artery anomalies, Cardiac anomalies, including aortic coarctation, and abnormalities of the Eyes. As part of a research project funded by the Gabriella Miller Kids First (GMKF) Pediatric Research Program, the genomes of 100 patients with PHACE syndrome were sequenced, identifying 15,752 rare, de novo variants. However, it is difficult to predict which of these variants contribute to this disorder since the large number of variants are spread over the entire genome.

The GMKF was created with the goal of expanding the understanding of structural birth defects, such as PHACE syndrome, and childhood cancer's underlying biology. Their Data Resource Portal (DRP) aggregates curated phenotype, disease, genetic data from clinical cohorts. Standardization of data using ontologies enables comparison of phenotype profiles across species and can aid in variant selection and prioritization, and ultimately, leverage the existing knowledge to uncover potential mechanisms of pathogenesis.

We curated phenotype profiles of 100 patients with PHACE Syndrome using the Human Phenotype Ontology (HPO) and the Mondo Disease Ontology. Forty-two curated phenotypic descriptions are currently available for this cohort and are freely available in the GMKF DRP (https://portal.kidsfirstdrc.org/).

Future work will leverage these curated patient phenotypic descriptions for cross-species comparison with phenotype data and their associated genotypes using computational algorithms from the Monarch Initiative. This cross-species comparison will help to identify candidate pathogenic genetic variants that are responsible for PHACE syndrome pathogenesis.

Variomes: a high recall search engine to support the curation of genetic variants

Emilie Pasche[1,2], Anaïs Mottaz[1,2], Déborah Caucheteur[1,2], Pierre-André Michel[1,2], Julien Gobeill[1,2] and Patrick Ruch[1,2]

[1] SIB Text Mining group, Swiss Institute of Bioinformatics, Geneva, Switzerland

[2] BiTeM group, Information Sciences, HES-SO / HEG, Carouge, Switzerland

ISB2021 session: session 4, October 5th

Category: Curation and Data Visualization Tools

Precision oncology relies on the use of treatments targeting specific genetic variants. However, the identification and interpretation of clinically actionable variants is a critical bottleneck. Searching for evidence in the literature is a labor-intensive task due to the lack of universally adopted standard nomenclature for variants. We thus propose a variant-specific literature search engine able to triage publications relevant to support an evidence-based decision. Our system searches within three pre-annotated document collections: MEDLINE abstracts, EuropePMC full-text articles and ClinicalTrials.gov. The system is based on two steps. First, we gather a large set of documents related to a particular case. To increase the recall, we are using query normalization and expansion. In particular, the variant expansion is performed with SynVar, a variant synonyms generator. Second, we apply different strategies to rank the publications, such as using the density of some specific named-entity types in the document. We assess the search effectiveness of the system using two experimental settings. Experimental setting 1: The literature retrieval task is tuned and evaluated using the TREC Precision Medicine 2018 and 2019 benchmarks consisting respectively in 50 and 40 topics. Almost two thirds (64%) of the publications returned in the top-five are relevant for clinical decision-support. Experimental setting 2: A comparison of Variomes with LitVar, a well-known search engine for genetic variants is performed, using 803 queries related to BRCA1 and BRCA2 variants. Variomes was able to retrieve on average 90.8% of the content, while LitVar retrieved on average 58.6%. To conclude, we are proposing here a competitive system to support the curation of variants for personalized oncology.

The clinical importance of tandem exon duplication-derived substitutions

Laura Martinez Gómez[1], Fernando Pozo[1], Thomas A. Walsh[1,2], Federico Abascal[3], Michael L. Tress[1]

[1] Bioinformatics Unit, Spanish National Cancer Research Centre (CNIO), C. Melchor Fernandez Almagro, 3, 28029 Madrid, Spain

[2] Eukaryotic Annotation Team, EMBL-EBI, Wellcome Genome Campus, Hinxton, Cambridgeshire CB10 1SA. UK

[3] Somatic Evolution Group, Wellcome Sanger Institute, Wellcome Genome Campus, Hinxton, Cambridgeshire CB10 1SA, UK

ISB2021 session: session 4, October 5th

Category: Comparative Genomics, Functional Annotation

Most coding genes in the human genome are annotated with multiple alternative transcripts. However, clear evidence for the functional relevance of the protein isoforms produced by these alternative transcripts is often hard to find. Alternative isoforms generated from tandem exon duplication derived substitutions are an exception. These splice events are rare, but have important functional consequences. Here, we have catalogued the 236 tandem exon duplication-derived substitutions annotated in the GENCODE human reference set. We find that more than 90% of the events have a last common ancestor in teleost fish, so are at least 425 million years old, and twenty-one can be traced back to the Bilateria clade. Alternative isoforms generated from tandem exon duplication-derived substitutions also have significantly more clinical impact than other alternative isoforms. Tandem exon duplication-derived substitutions have >25 times as many pathogenic and likely pathogenic mutations as other alternative events. Tandem exon duplication-derived substitutions appear to have vital functional roles in the cell and may have played a prominent part in metazoan evolution.

RepeatsDB in 2021: improved data and extended classification for protein tandem repeat structures

Alexander Miguel Monzon[1], Martina Bevilacqua[1], Damiano Clementel[1], Damiano Piovesan[1], Silvio C.E. Tosatto[1]

[1] Department of Biomedical Sciences, University of Padova, Padova, Italy

ISB2021 session: session 4, October 5th

Category: Curation and Data Visualization Tools

Almost twenty years ago the discovery of non-globular proteins has shaken the long-held structure-function paradigm where well-defined native protein structures are needed for function. The definition covers tandem repetitions, intrinsically disordered regions, aggregating domains and transmembrane domains. Tandem repeats proteins (TPRs) are composed of repetitions of these same or similar structural element modules. TPRs have been shown to act as an integral component of protein complexes and therefore to be involved in several biological functions, as well as neurodegenerative diseases.

TPRs are diverse in their amino acid sequences, structural states and functions. On one hand, data structures of solved repeat structures are accumulating, providing new possibilities for classification and detection. On the other hand there is an increasing need to organize and distribute specialized information on TPRs in an efficient way.

The RepeatsDB database (URL: https://repeatsdb.org/) provides manually curated annotations and classification for TPRs from the Protein Data Bank (PDB). The major conceptual change compared to the previous version is the hierarchical classification combining top levels based solely on structural similarity (Class > Topology > Fold) with two new levels (Clan > Family) requiring sequence similarity and describing repeat motifs in collaboration with Pfam. In particular, 'Clan', is a subfold that groups protein structures having a common sequence motif within the repeat (or part thereof). 'Family', will accommodate structures that have a common ancestor based on sequence similarity. Family classification aims at joining the sequence- and structure-based TR classifications of RepeatsDB and Pfam and to support the transfer of evolutionary and functional information.

Additionally, a new UniProt-centric view unifies the increasingly frequent annotation of structures from identical or similar sequences.

OBO Foundry in 2021

Randi Vita[1]

[1] Center for Infectious Disease and Vaccine Research, La Jolla Institute for Immunology, 9420 Athena Circle, La Jolla, CA, USA

ISB2021 session: session 4, October 5th

Category: Data Standards and Ontologies

The Open Biomedical Ontologies (OBO) Foundry (http://obofoundry.org) promotes sharing of biomedical ontologies and vocabularies, supporting collaborative ontology development and utilization of ontology design principles. There are over 200 biomedical ontologies in the OBO Library. The OBO Foundry is led by a collective of ontology developers who are committed to collaboration and adherence to shared principles. Collaborative development of ontologies using best practices and principles provides multiple advantages for biocurators. The OBO Foundry promotes consistent identifier use and metadata schemas for its resources and actively maintains a system of permanent URLs for those identifiers. This in turn enables resolution of ontology terms in a web browser for human curators, and provides a standard, computable way of referencing each term. To meet the requirements for inclusion in the OBO Foundry, ontologies must have a contact person and terms should have a clear label, definition, and source of definition (e.g. PubMed ID). The use of consistent identifiers ensures terms never disappear: even if they become obsolete, their identifiers remain valid, allowing curators to find the term corresponding to their annotation. The OBO Foundry's GitHub-backed website allows ontology editors to directly update the resource information, thus ensuring continual access to the latest information.

Developing ontologies that are easily re-usable and compatible with each other is a difficult process. The OBO Foundry principles facilitate this process by introducing shared conventions, such as standard identifiers, recommendations for open licenses, documentation, etc. In our growing community, it is hard to provide concrete development advice to implement these shared principles. This is especially true for new candidate ontologies, whose developers frequently seek advice on how to make their ontology more useful for others. To address this issue, the OBO Foundry developed a suite of tools, such as ROBOT, the Ontology Development Kit, and the OBO Dashboard, that make checking for errors (quality control) and conformance to OBO Foundry principles easier. The OBO Dashboard visually displays the results of OBO principle conformance checks and ROBOT reports, to help identify areas for improvement. The OBO Foundry actively encourages ontology developers to join the community, submit their ontologies for review, and participate in discussions on shaping the Foundry principles.

Extracting plant stress genes and functions from scientific literature to support annotation of switchgrass (*Panicum virgatum* L.) genes

Rita K. Hayford[1], Cecilia N. Arighi[1], Cathy H. Wu[1]
[1] Center for Bioinformatics and Computational Biology, University of Delaware Newark 19716

ISB2021 session: session 4, October 5th

Category: Functional Annotation

Plants are frequently exposed to a plethora of environmental stresses, affecting and limiting crop yield worldwide. These environmental stresses can be biotic (e.g., bacteria, viruses, fungi, and insects) or abiotic (e.g., drought, extreme temperatures, and salinity). Plants produce a wide variety of responses, for example, a change in the rate of photosynthesis and stomatal closure as they learn to endure these stresses. Crop improvement programs have focused on developing crops that can tolerate these stresses; thus, the number of publications on plant stress has significantly increased over the years. To help annotate Switchgrass genes involved in biotic and abiotic stress, we have established a pipeline that integrates text mining methods to efficiently retrieve plant stress genes from scientific literature and link them to their function. We used information from PudMed E-Utilities, Textpresso, pGenN, UniProt, and Europe PMC annotations for gene ontology and GeneRIF. The current collection, stored in MongoDB, includes 2,766 abstracts, 3,716 unique plant stress-responsive genes, 861 GO terms and 1,007 functional annotation GeneRIF sentences. We used this information to predict the role of stress- responsive genes in switchgrass, a bioenergy-relevant crop, using the collection of publications in this study to search for evidence described in other plants under stress. An additional outcome of this work is a set of publications for UniProt entries with annotations that can be submitted to the UniProt Knowledgebase via the community submission system. We expect that these annotated plant stress genes will be helpful for further understanding the mechanisms of stress tolerance in plants.

Integration of single-cell with bulk RNA-Seq to provide a unified view of gene expression and analysis tools in Bgee

Frederic Bastian[1,2]

[1] Department of Ecology and Evolution, University of Lausanne, 1015 Lausanne, Switzerland.

[2] SIB Swiss Institute of Bioinformatics, 1015 Lausanne, Switzerland.

ISB2021 session: session 4, October 5th

Category: Curation and Data Visualization Tools, Comparative Genomics, Data Standards and Ontologies

Single-cell RNA-Seq presents a new challenge for data integration in the life sciences, with not only large volumes of data but a large variety of samples and annotations. While many tools exist to analyse specific datasets, there is a lack of integration with the large amount of pre-existing information from bulk RNA-Seq as well as microarrays and other techniques. Challenges include annotation of data to relevant structures, from gross anatomy to cell types, curation and quality-control of datasets, and transformation of raw data into information which can be compared between data types.

Bgee (https://bgee.org/) is a database of gene expression expertise which already provides integration of gene expression information from bulk RNA-Seq, microarrays, ESTs and in situ hybridization, from diverse animal species, including human and model organisms. All data is curated to be healthy wild-type, and annotated to conditions of anatomy, development and aging, sex, and strain/population. Expression is integrated by computing for each gene and condition (i) calls of presence/absence of expression, and (ii) rank scores of expression. Both are comparable between data types, and allow to provide a Gene Page with a snapshot of gene expression over anatomy, TopAnat to compute over-represented conditions of expression for gene lists, Expression Comparison reporting homologous expression between species, and access to calls and processed data through R packages.

We present here the integration of single-cell RNA-Seq into Bgee release 15. We have defined curation and annotation criteria, methods to call expression present or absent, and ranks of expression over individual cells and cell populations. Bgee 15 provides all the functionalities of Bgee on the integration of single-cell with other expression information, seamlessly providing cell or organ-level information according to automatically computed relevance for each gene and view requested.

Rfam and the process of curating secondary structures of RNAs with 3D information

Nancy Ontiveros[1]

[1] European Molecular Biology Laboratory, European Bioinformatics Institute (EMBL-EBI), UK

ISB2021 session: session 4, October 5th

Category: Curation and Data Visualization Tools

Rfam (https://rfam.org) is a comprehensive database of RNA families, each represented by a manually curated alignment that includes a consensus secondary structure and a covariance model. Since its inception in 2002, Rfam has become the largest and most widely used database of non-coding RNA (ncRNA) families and currently includes 4,070 entries. The Rfam team is in the process of reviewing RNA families with available 3D information from Protein Data Bank (PDB). As of October 2021 the 3D structures of 122 Rfam families have been reported; however, many of these families were created before the structure was solved and are still based on the predicted secondary structures. Reviewing the families with 3D information allows Rfam to represent secondary structures more accurately. For example, the central part of the flavin mononucleotide (FMN) riboswitch is now organised by several additional base pairs and two pseudoknots. The first 20 updated families have been released and include riboswitches, coronavirus RNAs, spliceosomal RNAs, ribozymes, microRNAs, and other RNAs. To integrate the 3D structure information, the secondary structure annotations from RNA 3D structures are added to the alignments and then manually checked to ensure that the consensus structure is in agreement with the experimentally determined 3D information. As a result of the review process, we add or eliminate base pairs, include pseudoknot elements, and add other structural annotations, such as RNA 3D motifs (for example k-turn), RNA structural elements (stems, hairpin loops, junctions), and ligands (particularly in riboswitches). Rfam's aim is to continuously improve the quality of RNA families, and this targeted review of families using 3D information fills the gap between RNA predicted structures and the experimentally determined RNA 3D structures.