

*National Institute of Genetics International Symposium*

# Future perspectives of biological databases



- ⋮ Date & time: October 11, 2010. 13:00-15:45
- ⋮ Place: 11F AIST Waterfront Bio-IT Research Bldg.

<http://night.nig.ac.jp/labs/DnaData/>

Satellite meeting of Biocuration2010

<http://hinv.jp/biocuration2010/>



National Institute of Genetics International Symposium  
**"Future Perspectives of Biological Databases"**

11 October, 2010. 13:00-15:45  
11F AIST Tokyo Waterfront Bio-IT Research Building

**--- Program ---**

**Chairpersons:** Kazuho Ikeo and Yukiko Yamazaki (NIG, Japan)

Introduction to the NIG International Symposium

Kazuho Ikeo (National Institute of Genetics, Japan)

**SA1** Bioinformatics for Human Proteomics: Current State and Future Status

Amos Bairoch (Swiss Institute of Bioinformatics, Switzerland)

**SA2** Harnessing Wikipedia for Biocuration

Alex Bateman (Wellcome Trust Sanger Institute, UK)

**SA3** An Equation For a Vibrant Database: Curators + Journals + Community = Success

Tanya Berardini (Carnegie Institution for Science, USA)

**SA4** A Community of Biocurators

Pascale Gaudet (Northwestern University, USA)

**SA5** The Data Sharing Challenge. Learning Community from Consortia

Winston Hide (Harvard School of Public Health, USA)

**SA6** Coping with Increasingly Large Datasets - (Semi-)Automated Spatial Curation

Lorna Richardson (MRC Human Genetics Unit, UK)

**SA7** The Lessons Learned from Data Management for the Human Microbiome Project

Owen White (University of Maryland School of Medicine, USA)

**SA8** Future Outlook for Databases from the Perspective of Bioresources Databases

Yukiko Yamazaki (National Institute of Genetics, Japan)

**Commentators:**

Claire O'Donovan (EMBL-European Bioinformatics Institute, UK)

Tin Wee Tan (National University of Singapore, Singapore)

T. S. Keshava Prasad (Institute of Bioinformatics, India)

Steve Pettifer (The University of Manchester, UK)



## Abstracts:

### SA1

## **Bioinformatics for Human Proteomics: Current State and Future Status**

Amos Bairoch<sup>1</sup>

<sup>1</sup>*University of Geneva and Swiss Institute of Bioinformatics, Switzerland*

We are entering a new era in the exploration of the human proteome. Advances in technologies are allowing researchers to map with a much better resolution and improved accuracy this very complex universe. Such an exploration also requires a compendium of adequate bioinformatics tools, data repositories and knowledge resources. We are going to describe what are the state of the art in term of data and knowledge resources for human proteins: from repositories of proteomics data such as PRIDE, PeptideAtlas or Peptidome, databases such as SRMATlas and HPA and knowledgebases such as UniProtKB/Swiss-Prot. In September 2008, the UniProt/Swiss-Prot group achieved a major milestone: the first complete manual annotation of what is believed to be the full set of human proteins (derived from about 20'000 genes). This corpus of data is already quite rich in information pertinent to modern biomolecular medical research, but made us realize how large is the gap in our knowledge of human proteins in terms of functional information as well as protein characterization (PTMs, protein/protein interactions, subcellular locations, etc). This gap resides not only in the available experimental information, but also in the way this information has been stored, which is far from being sufficient to help researchers making sense of what all these human proteins do in our bodies! Therefore, in the framework of CALIPHO, a new interdisciplinary group created jointly by the University of Geneva and the Swiss Institute of Bioinformatics (SIB), we are developing neXtProt, a new human-centric protein knowledge resource. neXtProt is developed with the aim to help researchers answer pertinent questions relevant to human proteins. This requires answering 3 different challenges:

- 1) Add to the corpus of data on human proteins that is already in Swiss- Prot, a lot of additional information. We will import in neXtProt data originating from a variety of high-throughput approaches such as microarray, antibodies, proteomics, siRNAs, interactomics, etc. All of these data sets must be carefully selected so as to only provide high-quality data as we want to avoid creating a noisy and dirty compendium.
- 2) Organize the data in such a way that it is possible to make powerful queries in the most user-friendly environment. Here also, it is necessary to be able to capture the complexity and the heterogeneity of the data that will be available in neXtProt, yet make it easy for the user to forget this complexity!
- 3) Build a software platform that will allows tools ranging from sequence analysis to text and data mining to be integrated in various research environments so as to answer specific needs of academic and industrial users.

SA2

## **Harnessing Wikipedia for Biocuration**

Alex Bateman, Paul Gardner, Jennifer Daub

*Senior Investigator, Wellcome Trust Sanger Institute*

We have been using Wikipedia as a source of curated information for RNA families. I will present the latest status of this work and discuss more generally whether other databases might also use this approach.

### SA3

## **An Equation For a Vibrant Database: Curators + Journals + Community = Success**

Tanya Z. Berardini, Donghui Li, Raymond Chetty, Bob Muller, and Eva Huala

*The Arabidopsis Information Resource, Carnegie Institution for Science, Dept. of Plant Biology, Stanford, CA, USA*

Two years ago, The Arabidopsis Information Resource (TAIR) and the journal Plant Physiology began a collaboration to create an efficient mechanism for rapid and reliable transfer of the genetic and molecular data on Arabidopsis published in the journal into TAIR's public database. Since then, seven more plant journals have joined TAIR in the effort to involve the research community in direct data submission. TAIR now hosts a universal online data submission tool that allows authors with publications from any journal to submit their data directly to our curators. Once registered in the TAIR database, submitters can begin annotating as soon as they enter a DOI or PMID into the on-line form. The journals incorporate the URL for this form at critical points in the manuscript submission process. Hosting the submission tool ourselves allows us to streamline the data integration process and spend more time reviewing the annotations themselves instead of dealing with differences in journal-specific data formats. The TAIR-hosted form also makes the cost of collaborating with TAIR negligible for each journal which should allow for the rapid expansion of the set of publishers who promote direct data submission. Contributions from the community not only enrich the database and help to keep it current, but they also allow the database curators to focus on other publications that might otherwise not be read. All submissions are reviewed before integration into the database. This layer of review guarantees that the basic standards of annotation practiced at TAIR are applied to these data as well. The combination of all three communities biocurators, journal publishers and researchers working together ensure that the data available through TAIR are current and dynamic.

Journal participation and publicity have been instrumental in encouraging authors to submit their data to our community database. A small fraction of researchers contribute data without any prompting but they are very much the exception. We will present statistics on user submissions over the past two years and demonstrate the features of the online tool.

SA4

## **A community of biocurators**

Pascale Gaudet

*Northwestern University*

Biocuration is a difficult mission: biological data is highly heterogeneous, and there are few areas where guidelines for data preservation and exchange exist. This has hampered the speed, quantity and diversity of data possible to capture. Hence, a major challenge for the biocuration community is to explore innovative ways to capture, represent and display information. It is also essential to increase the involvement of researchers, publishers and funding agencies in this process. I will discuss the work of the International Society for Biocuration to build a community of biocurators to achieve these goals.

## SA5

### **The Data Sharing Challenge. Learning Community From Consortia**

Winston Hide<sup>1</sup>, Gabriel Altshuler, Oliver Hoffman<sup>2</sup> and Kimberly Begley<sup>2</sup>.

<sup>1</sup>*Department of Biostatistics, Harvard School of Public Health, USA*

<sup>2</sup>*Harvard School of Public Health Bioinformatics Core, USA*

Today's data generation and analysis environment is changing rapidly as the lure of technologies provide new opportunities to laboratory scientists. From physicians seeking rapid translation to the bedside, to environmentalists seeking new paradigms to understand the biome - large scale data generation is everywhere. How do we overcome the challenges of curation, management, interpretation and sharing of these data in the new multidisciplinary environment? One key approach is to leverage the expertise inherent in existing systems - bringing together tools and paradigms that work in subfields, and making them work together for the consortium. We will discuss the stem cell discovery engine. The approach brings together leading experts in stem cell biology, molecular profiles from stem cell systems, and seeks to combine data from diverse platforms under the unifying theme of common biology. We explore the use of gene lists and canonical pathways as a supervisor - and present pathway fingerprinting as a molecular approach to a broader problem. We explore how the community seeks to interact with its data - under the glare of reality.

## SA6

### **Coping with Increasingly Large Datasets - (Semi-)Automated Spatial Curation**

Lorna Richardson, Shanmugasundaram Venkataraman, Jeff Christiansen, Jianguo Rao, Peter Stevenson, Duncan Davidson and Richard Baldock

*MRC Human Genetics Unit, Institute of Genetics and Molecular Medicine, UK.*

Ultimately, the purpose of biological databases is to facilitate the use of the data they contain, to determine answers to biological questions. It stands to reason that this purpose is most likely to be achieved where the maximum amount of data can be accessed. It is with this in mind that EMAGE (E-Mouse Atlas of Gene Expression) a database of spatially integrated gene expression data in the developing mouse embryo, have pursued the inclusion of large datasets, notably from resources which are no longer supported, to ensure the continued availability of such valuable data. While for some time now we have used bioinformatics tools to assist in the curation of the meta-data pertaining to such large-scale screens, until recently it was only possible to carry out the spatial annotation of the expression data manually. This time-consuming process inevitably limited the quantity of data that could be included in EMAGE. The availability of a genome-wide screen containing ~19000 genes and ~360000 images rendered the manual annotation process unworkable, and it became necessary to develop a set of methods to automatically generate the spatial annotations. This undertaking required a considerable amount of effort and man-power, but the tools and methods developed will be applicable to any other consistently produced dataset, allowing further large depositions of data without the time limitations of manual annotation.

## SA7

### **The Lessons Learned From Data Management For The Human Microbiome Project**

Owen White<sup>1</sup>, Jennifer Wortman<sup>1</sup>

*<sup>1</sup>Institute for Genome Sciences, University of Maryland School of Medicine,  
Baltimore, MD*

The HMP is designed to fuel research into the microbes that live in the various environments of the human body. HMP data sets now include over hundreds of reference genomes isolated from the human body, as well as 16S ribosomal RNA, and whole metagenome shotgun sequencing of samples collected from multiple body sites and individuals. Successful utilization of this data requires the use of controlled vocabularies, the application of quality control measures, and the development of large-scale data management procedures. As part of this initiative, the HMP Data Analysis and Coordination center (DACC) has provided a data management and analysis infrastructure to support the collection, integration and standardization at several levels (see: [hmpdacc.org](http://hmpdacc.org)). As many people are aware the intersection of second generation sequencing technologies and the field of metagenomics is driving an explosion of data -- we will present our effort to meet these unique informatics challenges. A major goal of the HMP is to define a data set of metagenomic samples derived from healthy individuals to serve in comparisons between each sample as well as comparisons of data derived from individuals with disease phenotypes. We will provide an overview for how users may access this data, and the analysis tools that are now as well as tools that will be made available in the near future. We will also provide an overview of the metadata associated with the HMP samples, efforts to standardize the metadata across the HMP and international microbiome projects.

## **Future Outlook for Databases from the Perspective of Bioresources Databases**

Yukiko Yamazaki<sup>1</sup>

<sup>1</sup>*National Institute of Genetics, Japan*

Databases, like journals, are now an essential component of our research activities. However, the maintenance of databases and the timely, accurate, and complete incorporation into them of the wealth of data being created daily through scientific activities represent a difficult problem.

I have been working on bioresources databases and come to appreciate the need for closer links between journals and databases to improve the scientific value of databases. For example, it is frequently difficult to identify experimental materials from articles in journals. In experimental science, it is vital that one scientist's experiments are readily reproducible by others; this requires an environment that provides other scientists with ready access to identical experimental materials. This could become possible if journal requires a database address and/or an identifier of the material as a prerequisite of publication. There are also problems in the maintenance and management of gene lists. I believe that a dramatic improvement in this field would also be possible if database submission system were established in cooperation with journals, databases and communities.

In another regard, I have high expectations for ontology to make efficient use of databases. Bioresources databases deal with various biological species, such as animals, plants, and microorganisms. Differences in technical terms and in descriptions arising from differences in biological species can be overcome by means of ontology.

On the basis of my experiences in the field of bioresources databases, I will present my talk with a focus on the need for strong cooperation between databases and journals, and on my expectations for ontology.





Research Organization of Information and Systems

**National Institute of Genetics**