Biocuration 2010

The Conference of the International Society for Biocuration
The fourth International Biocuration Conference

11th-14th October, 2010
AIST Waterfront & Tokyo International Exchange Center,
Odaiba, Tokyo, Japan

Satellite Meetings

11th October, 2010 13:00-15:45 NIG International Symposium
“Future Perspectives of Biological Databases”

14th October, 2010 13:30-16:30 NIAS Symposium
“7th Rice Annotation Project Meeting (RAP7)”

http://hinv.jp/biocuration2010/
A breath of new life.

A catalyst of change.

To enable and usher in the new era.

HITACHI
Inspire the Next

www.hitachi.com
Hitachi, Ltd.
6-6, Marunouchi 1-chome, Chiyoda-ku, Tokyo, 100-8280 Japan  Phone:03-3258-1111
Biocuration 2010
The Conference of the International Society for Biocuration
& The 4th International Biocuration Conference

11-14 October 2010
AIST Tokyo Waterfront & Tokyo International Exchange Center
Odaiba, Tokyo

Organizing Committee

Chair:
Takashi Gojobori National Institute of Genetics

Vice-chair:
Tadashi Imanishi National Institute of Advanced Industrial Science and Technology

International members:
Amos Bairoch Swiss Institute of Bioinformatics/University of Geneva
Pascale Gaudet Northwestern University
Winston Hide Harvard School of Public Health
Minoru Kanehisa Kyoto University
Renate Kania HITS gGmbH/Scientific Databases and Visualization(SDBV)
Ilene Mizrachi NCBI/NLM/NIH
Barend Mons Netherlands Bioinformatics Centre
Haruki Nakamura Osaka University
Claire O'Donovan The European Bioinformatics Institute
Seung Yon Rhee Carnegie Institution for Science
Lorna Richardson EMAGE MRC Human Genetics Unit
Kazuki Saito RIKEN Plant Science Center/Chiba University
Toshihisa Takagi Research Organization of Information and Systems
Tetsuro Toyoda RIKEN(BASE)
Simon Twigger Medical College of Wisconsin
Owen White University of Maryland
Yukiko Yamazaki National Institute of Genetics

Local members:
Nobuyuki Fujita National Institute of Technology and Evaluation
Susumu Goto Kyoto University
Kazuho Ikeo National Institute of Genetics
Takeshi Itoh National Institute of Agrobiological Sciences
Hideya Kawaji RIKEN Omics Science Center
Akira Kinjo Osaka University
Chisato Yamasaki National Institute of Advanced Industrial Science and Technology

Supported by the Commemorative Organization for the Japan World Exposition(’70).
Biocuration2010 Abstract Book Editorial Team

Keiko Kondo
Takako Kuinaguchi
Akiko O. Noda
Yuichiro Hara
Miho Sera
# Table of Contents

Table of Contents ...................................................................................................................................... 3  
Sponsors.................................................................................................................................................... 5  
Welcome Message...................................................................................................................................... 6  
Floor Plan .................................................................................................................................................. 8  
Schedule ................................................................................................................................................... 12  
Conference Program ................................................................................................................................. 13  
Speaker Abstracts ...................................................................................................................................... 20  
  Plenary Talk1: **Dr. Minoru Kanehisa** .................................................................................................. 22  
  Plenary Talk2: **Dr. Paul D. Thomas** ............................................................................................... 23  
  Plenary Talk3: **Dr. Barend Mons** .................................................................................................... 24  
Session 1: **Large Scale, Automated, and Predictive Annotations** ......................................................... 25  
Session 2: **Outreach to User Communities & Community Annotation** ............................................... 27  
Session 3: **Biocuration in Metabolomics and Systems Biology** ................................................................ 30  
Session 4: **Asian Models of Biocuration-Related Activities** ................................................................ 33  
Session 5: **Data Integration Platforms and Data Sharing** ..................................................................... 36  
Session 6: **Future Directions and Challenges for Biocuration** ............................................................... 39  
Workshop 1: **Curation of Protein Structure and Chemicals** .................................................................. 42  
Workshop 2: **Curation Interface and Software** .................................................................................... 45  
Workshop 3: **Crop/Plant Genome Curation** ......................................................................................... 48  
Workshop 4: **New Approaches Towards Database Integration** ............................................................. 51  
Poster Abstracts .......................................................................................................................................... 54  
Exhibition .................................................................................................................................................. 106  
List of Participants ..................................................................................................................................... 107
Biocuration 2010 meeting is financially supported by the following grants.

- Commemorative Organization for the Japan World Exposition ‘70
- Kato Memorial Bioscience Foundation
- The Naito Foundation
- Sankyo Foundation of Life Science

Sponsors

- Ajinomoto Co., Inc.
- Bioinformatics Institute for Global Good, Inc. (BiGG)
- BioMed Central
- European Molecular Biology Organization
- The EMBO Meeting 2010
- The EMBO Meeting 2011
- FUJITSU LIMITED
- Hitachi, Ltd.
- Hitachi Government & Public Corporation System Engineering, Ltd.
- Hitachi Solutions, Ltd.
- The international ImMunoGeneTics information system (IMGT)
- Kirin Holdings Company, Limited
- Oxford University Press
- Yodosha

Exhibitors

- Database Center for Life Science, ROIS
- FUJITSU LIMITED
- Integrated Database and Systems Biology Team, BIRC, AIST
- Japan Biological Informatics Consortium (JBiC)
- Oxford University Press
- RIKEN Integrated Databases
Dear All,

I would like to warmly welcome all of you to join the Biocuration2010 meeting, the Conference of the International Society for Biocuration and the 4th International Biocuration Conference, here at Tokyo, Japan!! Indeed, it is our great pleasure to have all of you at Biocuration 2010, in the central region of the Tokyo metropolitan area.

The first Biocuration meeting was held in Asilomer, Pacific Grove, CA, in 2005. Since then, the idea of forming a society for curators and scientists working in the field of biocuration and bio-databases emerged and also presented at the 2007 International Biocurator meeting in San Jose, CA. At the International Biocurator Meeting in Berlin, Germany in the spring of 2009, the official inauguration was made that the International Society for Biocuration (ISB) has been established.

The goals of the International Society for Biocuration (ISB) are set up as follows: The Society helps define the profession of biocuration and its standpoint within the wider research community. It also provides a forum for biocurators, developers, researchers, and students to exchange experience and ideas. To ensure the stability of funding, the Society tries to promote the importance of this work. The Society also builds relationships with publishers and establishes bridges between researchers and databases through journal publishers. Moreover, the Society organizes a regular international meeting. Thus, the present meeting aims to invite all biocurators and their colleagues, providing them with a platform where all the participants including speakers and chairpersons can have intimate discussion and exchange of information under superb environments.

As we know, the acute advancements of biotechnologies such as next-gen sequencers lead us to production of a huge amount of biological and medical data. It naturally follows we have to meet immediate requirement of appropriate curation of the data for the database construction and the data analysis for biological importation and knowledge abstraction. So far in Japan, we keep administrating well-known databanks such as DNA Data Bank of Japan (DDBJ) and Protein Data Bank Japan (PDBj), Kyoto Encyclopedia of Genes and Genomes (KEGG) and the others. We also have established the annotated databases for several key organisms such as human, mouse, rice, etc. I can easily oversee that there are a large number of valuable databases submitted by Asian biocuration-related activities, showing the greater importance of this area as annotated data resources. Recognizing that the Biocuration2010 is the first Biocuration meeting in Asia, we would like to introduce these activities with emphasize on its significance.
We also encourage all the participants to enjoy staying in Japan and touching exquisite taste of Japanese culture. Odaiba, the meeting place, is one of the most popular regions in Japan where is close to the Narita International Airport as well as the metropolitan areas of central Tokyo. If you use a bullet train, Shinkansen, you have easy access to Kyoto and famous sightseeing spots.

Thank you again for all the participants to join this meeting, and we wish you a very happy stay in Japan. Let’s enjoy Biocuration!!

Best regards,

Takashi Gojobori
Chair, Biocuration 2010
The registration desk and cloak open at the first floor.

11 F

Conference Hall
(Plenary1 & NIG Symposium)

Headquarters

Speaker's Desk
(Slide Reception)

Elevator Hall

Coffee
Tokyo International Exchange Center

1F

Registration Desk

Cloak

Elevator Hall

National Museum of Emerging Science and Innovation

Administration Office, Tokyo International Exchange Center
Conference Program

[October 11th, Monday] AIST Tokyo Waterfront, Bio-IT Bldg. 11F

12:00 - 18:00 Registration desk opens at AIST Tokyo Waterfront, Bio-IT Bldg.

13:00 - 15:45

Satellite Meeting: National Institute of Genetics (NIG) International Symposium
"Future Perspectives of Biological Databases"
Chairpersons: Kazuho Ikeo (NIG, Japan) and Yukiko Yamazaki (NIG, Japan)

13:00 Introduction to the NIG International Symposium
Kazuho Ikeo (National Institute of Genetics, Japan)

SA1/13:05 Bioinformatics for Human Proteomics: Current State and Future Status
Amos Bairoch (Swiss Institute of Bioinformatics, Switzerland)

SA2/13:25 Harnessing Wikipedia for Biocuration
Alex Bateman (Wellcome Trust Sanger Institute, UK)

SA3/13:45 An Equation For a Vibrant Database: Curators + Journals + Community = Success
Tanya Berardini (Carnegie Institution for Science, USA)

SA4/14:05 A community of biocurators
Pascale Gaudet (Northwestern University, USA)

SA5/14:25 The Data Sharing Challenge. Learning Community From Consortia
Winston Hide (Harvard School of Public Health, USA)

SA6/14:45 Coping with Increasingly Large Datasets - (Semi-)Automated Spatial Curation
Lorna Richardson (MRC Human Genetics Unit, UK)

SA7/15:05 Lessons Learned from Data Management for the Human Microbiome Project
Owen White (University of Maryland School of Medicine, USA)

SA8/15:25 Closer Cooperation between Journals and Databases
Yukiko Yamazaki (National Institute of Genetics, Japan)

15:45 - 16:00 Coffee Break

16:00 - 16:30

Opening Remarks
Pascale Gaudet (Northwestern University, USA)
Takashi Gojobori (National Institute of Genetics, Japan)

16:30 - 17:30

Plenary Talk 1
PL1/16:30 Molecular Network Based Annotation of Genomes and Metagenomes
Minoru Kanehisa (Kyoto University, Bioinformatics Center, Japan)
<table>
<thead>
<tr>
<th>Time</th>
<th>Event</th>
</tr>
</thead>
<tbody>
<tr>
<td>18:00 - 20:00</td>
<td>Reception Restaurant <strong>LA TERRE</strong> in Miraikan (National Museum of Emerging Science and Innovation)</td>
</tr>
</tbody>
</table>

---

**October 12th, Tuesday**

**Tokyo International Exchange Center**

- **8:30 - 17:00**
  - Registration desk opens at Tokyo International Exchange Center

- **9:30 - 10:30**
  - **Plenary Talk 2**
    - **PL2/9:30**
      - *Evolutionary Relationships as a Paradigm for Integrating Biological Knowledge*
      - Paul D. Thomas (University of Southern California, USA)

- **10:30 - 12:30**
  - **Session 1: Large Scale, Automated, and Predictive Annotations**
    - Chairpersons: Claire O'Donovan (EBI) and Barend Mons (Netherlands Bioinformatics Centre)
    - **S1-1/10:30**
      - *Comparison of Computationally- and Manually-Assigned Gene Ontology Annotations to Improve Functional Characterization of Gene Products*
      - Maria C. Costanzo (Stanford University, USA)
    - **S1-2/11:00**
      - *Automatic Protein Clustering as a Basis of Automatic Annotation*
      - Naoki Sato (The University of Tokyo, Japan)
    - **S1-3/11:30**
      - *Transcriptome in a Dynamic System with Next Gen Sequencers*
      - Hideya Kawaji (RIKEN Omics Science Center, Japan)
    - **S1-4/12:00**
      - *UniRule - Automatic Annotation In UniProtKB*
      - Alan Bridge (SIB, Switzerland)

- **12:30 - 13:30**
  - Lunch & Poster presenters place posters on boards

- **13:30 - 15:30**
  - **Session 2: Outreach to User Communities & Community Annotation**
    - Chairpersons: Maria Costanzo (Stanford University), Mike Cherry (Stanford University) and Winston Hide (Harvard School of Public Health)
    - **S2-1/13:30**
      - *Community Annotation in Biology*
      - Raja Mazumder (PIR, Georgetown University Medical Center, USA)
    - **S2-2/14:00**
      - *Nanopublication by Community Annotation in the ConceptWiki*
      - Christine Chichester (Netherlands Bioinformatics Centre, Netherlands)
    - **S2-3/14:15**
      - *The Challenge of Eukaryote Genome Annotation At Genoscope*
      - Betina Porcel (Genoscope, France)
    - **S2-4/14:30**
      - *The Neuroscience Information Framework: A Unified Semantic Framework for Discovery and Integration of Biomedical Data and Resources on the Web*
      - Jeffrey Grethe (UCSD, USA)
    - **S2-5/14:45**
      - *Incorporating Community Annotation Interfaces into the CIPRO2.5 Database with Comprehensible Sketches to Support Quick Annotations of Proteome Data*
      - Keisuke Ueno (Hokkaido University, Japan)

- **15:00**
  - Panel Discussion: *Outreach to User Communities - Who are they and what do they need?*
15:30 - 15:45 Coffee Break

15:45 - 17:15 Concurrent Workshops 1 & 2

### Workshop 1: Curation of Protein Structure and Chemicals (Conference Room 1@4F)

**Chairpersons:** Susumu Goto (Kyoto Univ, Japan) and Akira R. Kinjo (Osaka Univ, Japan)

<table>
<thead>
<tr>
<th>Time</th>
<th>Title</th>
<th>Speaker</th>
</tr>
</thead>
<tbody>
<tr>
<td>W1-1/15:45</td>
<td><strong>Comparative Protein Structure Modeling: An Effective Means to Explore Protein Function</strong></td>
<td>Česlovas Venclovas (Inst Biotech, Lithuania)</td>
</tr>
<tr>
<td>W1-2/16:15</td>
<td><strong>Biomacromolecular Structure Annotation</strong></td>
<td>Jasmine Young (RCSB PDB, Rutgers University, USA)</td>
</tr>
<tr>
<td>W1-3/16:30</td>
<td><strong>ChEBI, an Open-access Chemistry Resource for the Life Sciences: Facilities for On-line Submission and Curation</strong></td>
<td>Marcus Ennis (EBI, UK)</td>
</tr>
<tr>
<td>W1-4/16:45</td>
<td><strong>Functional Annotation of Sugar Databases</strong></td>
<td>Frederique Lisacek (SIB, Switzerland)</td>
</tr>
<tr>
<td>W1-5/17:00</td>
<td><strong>Data Standardization for Glyco-informatics</strong></td>
<td>Kiyoko Aoki-Kinoshita (Soka University, Japan)</td>
</tr>
</tbody>
</table>

### Workshop 2: Curation Interface and Software (International Conference Hall@3F)

**Chairpersons:** Xiaodong Wang (WormBase, CalTech, USA) and Kimberly Van Auken (WormBase, CalTech, USA)

<table>
<thead>
<tr>
<th>Time</th>
<th>Title</th>
<th>Speaker</th>
</tr>
</thead>
<tbody>
<tr>
<td>W2-1/15:45</td>
<td><strong>Curation of Protein Domain Models for the Conserved Domain Database (CDD)</strong></td>
<td>Aron Marchler-Bauer (NCBI, USA)</td>
</tr>
<tr>
<td>W2-2/16:00</td>
<td><strong>InterPro Curation: Integrating Predictive Protein Signatures into Biological Hierarchies</strong></td>
<td>Sarah Burge (EMBL-EBI, UK)</td>
</tr>
<tr>
<td>W2-3/16:15</td>
<td><strong>Manual Biocuration, a UniProtKB/Swiss-Prot Perspective</strong></td>
<td>Sylvain Poux (SIB, Switzerland)</td>
</tr>
<tr>
<td>W2-4/16:30</td>
<td><strong>ISA Software Suite: Supporting Standards-Compliant Experimental Annotation and Enabling Curation at the Community Level</strong></td>
<td>Philippe Rocca-Serra (University of Oxford, OeRC, UK)</td>
</tr>
<tr>
<td>W2-5/16:45</td>
<td><strong>Ontology-based Tools to Enhance the Curation Workflow</strong></td>
<td>Patricia Whetzel (Stanford University, USA)</td>
</tr>
<tr>
<td>W2-6/17:00</td>
<td><strong>Gene Curation Software at the Rat Genome Database</strong></td>
<td>Stan Laulederkind (Medical College of Wisconsin, USA)</td>
</tr>
</tbody>
</table>

17:30 - 18:30

**Poster Presentations (odd number)**

18:30 - 19:30

**Poster Presentations (even number)**
8:30 - 17:00 Registration desk opens at Tokyo International Exchange Center

9:00 - 10:00

**Plenary Talk 3**

PL3/9:00  *A Few Hundred Minds, a Million Minds or Millions of Minds?*

Barend Mons (Netherlands Bioinformatics Centre, Netherlands)

10:00 - 12:00

**Session 3: Biocuration in Metabolomics and Systems Biology**

Chairpersons: Kazuki Saito (RIKEN) and Seung Yon Rhee (Carnegie Institution for Science, USA)

S3-1/10:00  *Combining Computational Prediction and Manual Curation to Create Plant Metabolic Pathway Databases*

Peifen Zhang (Carnegie Institution for Science, USA)

S3-2/10:30  *KaPPA-View4: A Pathway Database for Gene Co-expression and Metabolite Co-accumulation Analysis*

Nozomu Sakurai (Kazusa DNA Research Institute, Japan)

S3-3/10:55  *MassBank: A Reference Database for Chemical Identification of Metabolites Detected by Mass Spectrometry*

Takaaki Nishioka (Keio University, Japan)

S3-4/11:20  *Plant Metabolic Data Curation and Its Integration into Solanaceae Genomic Databases*

Anuradha Pujar (Boyce Thompson Institute for Plant Research, USA)

S3-5/11:40  *The EcoCyc Database - Integrating and Transferring Knowledge about E. coli*

Ingrid Keseler (SRI International, USA)

12:00 - 12:15  Group Photo

12:15 - 13:30  Lunch

13:30 - 15:30

**Session 4: Asian Models of Biocuration-Related Activities**

Chairpersons: Takashi Gojobori (NIG)

S4-1/13:30  *Curation for Systems Biology Resources in an Academic Model*

Keshava Prasad (Institute of Bioinformatics, India)

S4-2/13:55  *Identification Of Causative Variations Based On The Technology Of Next Generation Sequencing*

Yixue Li (Shanghai Center for Bioinformation Technology, China)

S4-3/14:20  *Asian Models of Micorbial Resource Information Network*

Juncai Ma (Institute of Microbiology, CAS, China)

S4-4/14:45  *Biocuration of Genomic Data of Non-model Species in Taiwan*

Tzen-Yuh Chiang (National Cheng Kung University, Taiwan)

S4-5/15:10  *Data Integration Model and GUIs Used in Human Genome Network Platform*

Kazuho Ikeo (National Institute of Genetics, Japan)
<table>
<thead>
<tr>
<th>Time</th>
<th>Event</th>
</tr>
</thead>
<tbody>
<tr>
<td>15:30 - 15:45</td>
<td>Coffee Break</td>
</tr>
<tr>
<td>15:45 - 17:15</td>
<td><strong>Concurrent Workshops 3 &amp; 4</strong></td>
</tr>
<tr>
<td></td>
<td><strong>Workshop 3: Crop/Plant Genome Curation (Conference Room 1@4F)</strong></td>
</tr>
<tr>
<td></td>
<td>Chairpersons: Takeshi Itoh (NIAS, Japan) and Mary L. Schaeffer (USDA-ARS, USA)</td>
</tr>
<tr>
<td>W3-1/15:45</td>
<td><em>Data Curation Infrastructure in the iPlant Discovery Environment</em></td>
</tr>
<tr>
<td></td>
<td>Chris Jordan (The University of Texas at Austin, USA)</td>
</tr>
<tr>
<td>W3-2/16:15</td>
<td><em>Application of Next-Generation Sequence Data For Rice Genome Analyses</em></td>
</tr>
<tr>
<td></td>
<td>Yoshihiro Kawahara (NIAS, Japan)</td>
</tr>
<tr>
<td>W3-3/16:30</td>
<td><em>MAIZEGDB.ORG, the Maize Genetics Cooperation and the 2500 MB B73</em></td>
</tr>
<tr>
<td></td>
<td>Genome-Generated Tsunami</td>
</tr>
<tr>
<td></td>
<td>Mary L. Schaeffer (USDA-ARS, USA)</td>
</tr>
<tr>
<td>W3-4/16:45</td>
<td><em>Arabidopsis thaliana: Further Exploiting This Plant Reference Genome</em></td>
</tr>
<tr>
<td></td>
<td>Mara Sangiovanni (University of Naples Federico II, Italy)</td>
</tr>
<tr>
<td>W3-5/17:00</td>
<td><em>A New QTL Curation Approach, Analysis and Linking QTLs to Genomes at SGN</em></td>
</tr>
<tr>
<td></td>
<td>Isaak Tecle (Cornell University, USA)</td>
</tr>
<tr>
<td></td>
<td><strong>Workshop 4: New Approaches Towards Database Integration (International Conference Hall@3F)</strong></td>
</tr>
<tr>
<td></td>
<td>Chairpersons: Chisato Yamasaki (BIRC, AIST, Japan) and Simon Twigger (Medical College of Wisconsin, USA)</td>
</tr>
<tr>
<td>W4-1/15:45</td>
<td><em>Hyperlink Management System for Creating Maintenance-Free Hyperlinks among Major Biological Databases</em></td>
</tr>
<tr>
<td></td>
<td>Tadashi Imanishi (BIRC, AIST, Japan)</td>
</tr>
<tr>
<td>W4-2/16:00</td>
<td><em>Semantic Encoding of Complex Information</em></td>
</tr>
<tr>
<td></td>
<td>Steven Vercruysse (Norwegian University of Science and Technology, Norway)</td>
</tr>
<tr>
<td>W4-3/16:15</td>
<td><em>NamesforLife Semantic Resolution Services for the Life Sciences: Moving towards an Extensible and Interoperable System for Naming</em></td>
</tr>
<tr>
<td></td>
<td>George M. Garrity (NamesforLife, LLC and Michigan State University, USA)</td>
</tr>
<tr>
<td>W4-4/16:30</td>
<td><em>Bio2RDF: Convert, Provide and Reuse</em></td>
</tr>
<tr>
<td></td>
<td>Marc-Alexandre Nolin (Laval University, Canada)</td>
</tr>
<tr>
<td>W4-5/16:45</td>
<td><em>RIKEN SciNetS: A Cloud of World's First Publication Medium of Databases on the Semantic Web</em></td>
</tr>
<tr>
<td></td>
<td>Tetsuro Toyoda (RIKEN BASE, Japan)</td>
</tr>
<tr>
<td>W4-6/17:00</td>
<td><em>TogoDB + TogoWS: a Data Integration Platform for the Semantic Web</em></td>
</tr>
<tr>
<td></td>
<td>Toshiaki Katayama (Univ Tokyo, Japan) and Mitsuteru Nakao (DBCLS, Japan)</td>
</tr>
<tr>
<td>17:30 - 18:30</td>
<td><strong>Poster Presentations (even number)</strong></td>
</tr>
<tr>
<td>18:30 - 19:30</td>
<td><strong>Poster Presentations (odd number)</strong></td>
</tr>
</tbody>
</table>
8:30 - 13:00 Registration desk opens at Tokyo International Exchange Center

9:00 - 10:30

**Session 5 : Data Integration Platforms and Data Sharing**
Chairpersons : Amos Bairoch (SIB) and Pascale Gaudet (Northwestern University)

<table>
<thead>
<tr>
<th>Time</th>
<th>Title</th>
<th>Speaker</th>
</tr>
</thead>
<tbody>
<tr>
<td>9:00</td>
<td>Omics Data Sharing</td>
<td>Susanna Sansone (University of Oxford, OeRC, UK)</td>
</tr>
<tr>
<td>9:30</td>
<td>H-InvDB: A Comprehensive Annotation Resource For Human Transcriptome</td>
<td>Chisato Yamasaki (BIRC, AIST, Japan)</td>
</tr>
<tr>
<td>9:45</td>
<td>Literature Curation of Protein Interactions: Measuring Agreement Across Major Public Databases</td>
<td>Andrei Turinsky (Hospital for Sick Children, Canada)</td>
</tr>
<tr>
<td>10:00</td>
<td>neXtProt, a New Human-Centric Protein Knowledge Resource</td>
<td>Lydie Lane (University of Geneva, Switzerland)</td>
</tr>
<tr>
<td>10:15</td>
<td>UniProt Knowledgebase: A Hub of Integrated Protein Data</td>
<td>Michele Magrane (European Bioinformatics Institute, UK)</td>
</tr>
</tbody>
</table>

10:30 - 10:45 Coffee Break

10:45 - 12:20

**Session 6 : Future Directions and Challenges for Biocuration**
Chairpersons : Renate Kania (HITS gGmbH) and Tadashi Imanishi (BIRC, AIST)

<table>
<thead>
<tr>
<th>Time</th>
<th>Title</th>
<th>Speaker</th>
</tr>
</thead>
<tbody>
<tr>
<td>10:45</td>
<td>MetaCuration Standards and Minimum Information about a Bioinformatics Investigation</td>
<td>Tin Wee Tan (National University of Singapore)</td>
</tr>
<tr>
<td>11:10</td>
<td>Utopia Documents: Linking Literature and Data</td>
<td>Steve Pettifer (The University of Manchester, UK)</td>
</tr>
<tr>
<td>11:35</td>
<td>AMIS, the Article Minimum Information Standard</td>
<td>Delphine Dauga (Inst Dev Biol Marseille-Luminy, France)</td>
</tr>
<tr>
<td>11:50</td>
<td>Publishing Interactive Articles: Integrating Journals and Biological Databases</td>
<td>Karen Yook (WormBase Curator, USA)</td>
</tr>
<tr>
<td>12:05</td>
<td>The SciKnowMine System: Infrastructure for Biocuration Applied to MGI Document Triage</td>
<td>Gully Burns (USC/ISI, USA)</td>
</tr>
</tbody>
</table>

12:20 - 12:30

**Closing Remark**
Tadashi Imanishi (BIRC, AIST, Japan)

12:30 - 13:00 Poster presenters remove posters from boards
12:30 - 13:30 Lunch
Satellite Meeting: NIAS Symposium "7th Rice Annotation Project Meeting"

13:30 - 16:30

13:30 Opening Remarks
   Takuji Sasaki (National Institute of Agrobiological Sciences, Japan)

13:35 Chromosomal Genomics: Unlocking the 5GB Barley Genome
   Klaus Mayer (MIPS, German Research Center for Environmental Health, Germany)

14:15 The Rice Annotation Project Database for Comparative Genomics
   Tsuyoshi Tanaka (National Institute of Agrobiological Sciences, Japan)

14:35 Maize Genome Annotation: Transitioning from the Sequencing Project to MaizeGDB
   Mary Schaeffer (MaizeGDB USDA/ARS Plant Genetics Research Unit, USA)

14:55-15:05 Break

15:05 DDBJ Read Annotation Pipeline: A Cloud Computing Based Analytical Tool for Next-generation Sequencing Data
   Yasukazu Nakamura (National Institute of Genetics, Japan)

15:45 Annotation of Plant Proteins in UniProtKB/Swiss-Prot
   Damien Lieberherr (Swiss Institute of Bioinformatics, Switzerland)

16:05 Characterizing Genetic Diversity and Creating Novel Gene Pools in Rice for Trait Dissection and Gene Function Discovery
   Ramil P. Mauleon (International Rice Research Institute, Philippines)

18:00 - 21:00 Excursion: Tokyo Bay Dinner Cruise by “Yakatabune”

* Registration desk will be open during the following hours:
  October 11 (Mon): 12:00-18:00 at AIST Bio-IT Bldg.
  October 12 (Tue): 8:30-17:00 at Tokyo International Exchange Center
  October 13 (Wed): 8:30-17:00 at Tokyo International Exchange Center
  October 14 (Thu): 8:30-13:00 at Tokyo International Exchange Center
Speaker Abstracts

Plenary Talk 1: Molecular Network Based Annotation of Genomes and Metagenomes
Plenary Talk 2: Evolutionary Relationships as a Paradigm for Integrating Biological Knowledge
Plenary Talk 3: A Few Hundred Minds, a Million Minds or Millions of Minds?

Session 1: Large Scale, Automated, and Predictive Annotations
S1-1 Comparison of Computationally- and Manually-Assigned Gene Ontology Annotations to Improve Functional Characterization of Gene Products
S1-2 Automatic Protein Clustering as a Basis of Automatic Annotation
S1-3 Transcriptome in a Dynamic System with Next Gen Sequencers
S1-4 UniRule - Automatic Annotation In UniProtKB

Session 2: Outreach to User Communities & Community Annotation
S2-1 Community Annotation in Biology
S2-2 Nanopublication by Community Annotation in the ConceptWiki
S2-3 The Challenge of Eukaryote Genome Annotation At Genoscope
S2-4 The Neuroscience Information Framework: A Unified Semantic Framework for Discovery and Integration of Biomedical Data and Resources on the Web
S2-5 Incorporating Community Annotation Interfaces into the CIPRO2.5 Database with Comprehensible Sketches to Support Quick Annotations of Proteome Data

Session 3: Biocuration in Metabolomics and Systems Biology
S3-1 Combining Computational Prediction and Manual Curation to Create Plant Metabolic Pathway Databases
S3-2 KaPPA-View4: A Pathway Database for Gene Co-expression and Metabolite Co-accumulation Analysis
S3-3 MassBank: A Reference Database for Chemical Identification of Metabolites Detected by Mass Spectrometry
S3-4 Plant Metabolic Data Curation and Its Integration into Solanaceae Genomic Databases
S3-5 The EcoCyc Database - Integrating and Transferring Knowledge about E. coli

Session 4: Asian Models of Biocuration-Related Activities
S4-1 Curation for Systems Biology Resources in an Academic Model
S4-2 Identification Of Causative Variations Based On The Technology Of Next Generation Sequencing
S4-3 Asian Models of Micorbial Resource Information Network
S4-4 Biocuration of Genomic Data of Non-model Species in Taiwan
S4-5 Data Integration Model and GUIs Used in Human Genome Network Platform

Session 5: Data Integration Platforms and Data Sharing
S5-1 Omics Data Sharing
S5-2 H-InvDB: A Comprehensive Annotation Resource For Human Transcriptome
S5-3 Literature Curation of Protein Interactions: Measuring Agreement Across Major Public Databases
S5-4 neXtProt, a New Human-Centric Protein Knowledge Resource
Session 6: Future Directions and Challenges for Biocuration

S6-1 MetaCuration Standards and Minimum Information about a Bioinformatics Investigation
S6-2 Utopia Documents: Linking Literature and Data
S6-3 AMIS, the Article Minimum Information Standard
S6-4 Publishing Interactive Articles: Integrating Journals and Biological Databases
S6-5 The SciKnowMine System: Infrastructure for Biocuration Applied to MGI Document Triage

Workshop 1: Curation of Protein Structure and Chemicals (Conference Room 1@4F)

W1-1 Comparative Protein Structure Modeling: An Effective Means to Explore Protein Function
W1-2 Biomacromolecular Structure Annotation
W1-3 ChEBI, an Open-access Chemistry Resource for the Life Sciences: Facilities for On-line Submission and Curation
W1-4 Functional Annotation of Sugar Databases
W1-5 Data Standardization for Glyco-informatics

Workshop 2: Curation Interface and Software (International Conference Hall@3F)

W2-1 Curation of Protein Domain Models for the Conserved Domain Database (CDD)
W2-2 InterPro Curation: Integrating Predictive Protein Signatures into Biological Hierarchies
W2-3 Manual Biocuration, a UniProtKB/Swiss-Prot Perspective
W2-4 ISA Software Suite: Supporting Standards-Compliant Experimental Annotation and Enabling Curation at the Community Level
W2-5 Ontology-based Tools to Enhance the Curation Workflow
W2-6 Gene Curation Software at the Rat Genome Database

Workshop 3: Crop/Plant Genome Curation (Conference Room 1@4F)

W3-1 Data Curation Infrastructure in the iPlant Discovery Environment
W3-2 Application of Next-Generation Sequence Data For Rice Genome Analyses
W3-3 MAIZEGDB.ORG, the Maize Genetics Cooperation and the 2500 MB B73 Genome-Generated Tsunami
W3-4 Arabidopsis thaliana: Further Exploiting This Plant Reference Genome
W3-5 A New QTL Curation Approach, Analysis and Linking QTLs to Genomes at SGN

Workshop 4: New Approaches Towards Database Integration (International Conference Hall@3F)

W4-1 Hyperlink Management System for Creating Maintenance-Free Hyperlinks among Major Biological Databases
W4-2 Semantic Encoding of Complex Information
W4-3 NamesforLife Semantic Resolution Services for the Life Sciences: Moving towards an Extensible and Interoperable System for Naming
W4-4 Bio2RDF: Convert, Provide and Reuse
W4-5 RIKEN SciNetS: A Cloud of World's First Publication Medium of Databases on the Semantic Web
W4-6 TogoDB + TogoWS: a Data Integration Platform for the Semantic Web
The large-scale molecular datasets generated by genome sequencing and other high-throughput experimental technologies are the basis for understanding life as a molecular system and for developing practical applications in medical, pharmaceutical, and environmental sciences. The key to linking such large-scale datasets to practical values lies in bioinformatics technologies, not only in terms of computational methods, but also in terms of knowledge bases. We have been developing KEGG (http://www.genome.jp/kegg/), a database resource that integrates genomic, chemical, and systemic functional information. In particular, gene catalogs in the completely sequenced genomes are linked to molecular networks (pathways, etc.) representing higher-level systemic functions of the cell, the organism, and the ecosystem. Major efforts have been undertaken to manually create a knowledge base for such systemic functions by capturing and summarizing experimental knowledge in computable forms, especially in the form of molecular network diagrams called KEGG pathway maps. Other forms have also been used to supplement KEGG pathway maps, including hierarchical lists for BRITE functional hierarchies (ontologies) and simple lists for membership information in the KEGG MODULE and KEGG DISEASE databases. The linkage between genes and proteins to KEGG pathway nodes and BRITE hierarchy nodes is made through the KEGG Orthology (KO) system, a collection of ortholog groups identified by the K numbers. The KEGG annotation is essentially cross-species annotation giving K numbers to orthologous genes in all available genomes, and is currently done as follows. (i) Experimental evidence on known functions is organized in the KO database, which is created together with the KEGG PATHWAY and KEGG BRITE databases. (ii) Gene catalogs of complete genomes are generated from RefSeq and other public resources. (iii) All pairs of genomes (gene catalogs) are compared by the SSEARCH program, and the GFIT tables are generated showing for each gene in a genome the information about best-hit genes in all other genomes. (iv) GFIT tables are continuously updated, and the automatic version of the KOALA tool presents to human annotators a summary of discrepancies between its K number assignment and the current annotation. (v) Discrepancies are examined by annotators with the manual version of KOALA and other tools. (vi) Annotation results are mapped to KEGG pathways and BRITE hierarchies for inferring systemic functions of individual organisms, groups of organisms (e.g., pangenomes), and combinations of organisms (e.g., host-pathogen and human-microbiome relationships). In this talk I will first give an overview of the KEGG annotation procedures, and then present our new efforts to integrate disease and drug information into the KEGG molecular networks.
Evolutionary Relationships as a Paradigm for Integrating Biological Knowledge

Paul D. Thomas, PhD

Associate Professor and Director Division of Bioinformatics Department of Preventive Medicine USC Keck School of Medicine

The common ancestry of all living organisms means that discoveries in any one biological model organism may shed light on the biology of even distant relatives. Thus, the evolutionary relationships, or natural classification, can be used to integrate the knowledge obtained across disparate organisms. I discuss examples of how evolutionary considerations are guiding curation—or, more properly, “integrative biology”—projects addressing both closely and distantly related groups of organisms, from strains of “E. coli” to species spanning the tree of life.
Biocuration has encountered the problems of dealing with heterogeneous data and making it interoperable while ensuring that annotation, which includes the recognition-and-reward system of scientific publishing, fits into a seamless workflow. The early attempts to involve the wider scientific community in the annotation and curation process have also met with serious setbacks. However, it is inevitable that the broad community must become involved in annotation. The Concept Web Alliance partners have started an ambitious collaboration based on lessons learned and recent technical and social developments. Our ultimate goal is to create a sustainable future for a large-scale, community editable store of disambiguated scientific assertions, exemplifying a new paradigm in life sciences data accumulation. By removing the barriers of differing data formats and harnessing a distributed community-based concept annotation, the scientific community can expect a significant increase in productivity. Even more important, the semantic enrichment of this unified data resource will open the door to sophisticated approaches to reasoning, creating new knowledge that cannot be obtained using more traditional means. We will draw from the mental resources of an extended scientific community in an innovative and complex, yet ‘daily practice’, manner that promises a profound impact on our ability to use existing data to generate new knowledge with the maximum conceivable serendipity.
SI-1

Comparison of computationally- and manually-assigned Gene Ontology annotations to improve functional characterization of gene products

Maria C. Costanzo, Rama Balakrishnan, Karen R. Christie, Eunie L. Hong, Julie Park, J. Michael Cherry, and The Saccharomyces Genome Database Project.

Department of Genetics, Stanford University, Palo Alto, CA, USA

The Gene Ontology (GO) describes molecular functions, biological processes, and cellular components of gene products using controlled-vocabulary terms that are related to each other in a structure that facilitates computing on GO annotations within and across species. Experimentally-based GO annotations that are manually curated from the literature are often used to predict the functions of related uncharacterized proteins. The accuracy of such annotations is thus critically important, particularly for a well-studied model organism such as Saccharomyces cerevisiae which is frequently used as the source of the experimental data.

Comparison of experimentally-based annotations with those predicted by computational methods for the same gene products may reveal inaccuracies in curation of the experimental data, and could additionally be used to evaluate and improve the computational methods. We will present the results of an analysis at SGD that identified four major reasons for discrepancies between the two kinds of annotation. Some discrepancies revealed cases in which human error led to errors or omissions in the manual curation, prompting prioritization for review and correction. In another category, the computational annotations were not supported or were refuted by the literature, thereby suggesting ways in which the accuracy of the prediction methods could be improved. Yet another type of discrepancy resulted from issues with the GO structure, such as missing parentage for certain terms, leading to reexamination and improvement of the ontology. Finally, some discrepancies arose because the computational predictions were entirely novel, and no relevant experimental evidence was available. These cases highlight potential interesting new avenues for experimentation.

SI-2

Automatic protein clustering as a basis of automatic annotation

Naoki Sato

Graduate School of Arts and Sciences, University of Tokyo, Japan

Development of new generation sequencers enabled genome sequencing feasible for every organism in a laboratory. A typical data flow of de novo sequencing includes (1) assembly of sequence reads, (2) estimation of open reading frames, (3) annotation of proteins, and (4) finding RNA genes. The annotation is normally performed by BLASTP searches against several different databases. However, it is usually hard to find a plausible annotation by just looking at the results of BLASTP searches.

Here I propose a potentially automatic method of annotation that exploits automatic protein clustering using the software GCLUST, which estimates proper similarity threshold for each list of homologs using 'entropy-optimized organism count' method (Sato 2009). The software has been used to construct a homolog database including both prokaryotic and eukaryotic proteins (http://gclust.c.u-tokyo.ac.jp/). For use in genome annotation, we need de novo clustering including many genomes of related organisms as well as genomes of representative organisms. Application of protein clustering in the annotation in Arthrospira platensis was the first successful case (Nariikawa et al. 2010). I present here results of protein clustering of total predicted proteins in two draft genomes of cyanobacteria along with total predicted proteins of 41 cyanobacteria available at NCBI. For each of the resultant protein clusters, an alignment and a phylogenetic tree were also prepared for assistance in functional annotation. The quality of alignments was evaluated by counting ill-aligned proteins (missing N- or C-terminus, or insertion/deletion), which was 4-13% of total predicted proteins in most cyanobacterial genomes. Annotation may be automated by extracting significant key words already assigned for member proteins of clusters or by comparison with reference protein clusters.
Identification And Activity Profiling Of Promoters With Next Generation Sequencers

Hideya Kawaji
RIKEN Omics Science Center, Japan

Genome-wide gene expression analyses have been possible with microarray platform only after identification of RNA structures, and such expression analyses have been separated from the genome annotation of mRNAs. However, with the remarkable progress of the sequencing technologies, the identification of RNA’s (partial) structures and the quantification of their abundances can be achieved at once, which leads to the identification of novel structure existing only in the profiled samples as well as conventional gene expression. A wide range of experimental protocols, such as RNA-seq sequencing randomly fragmented mRNA molecules, small RNA-seq sequencing microRNA-size RNAs, CAGE (Cap Analysis Gene Expression) sequencing 5'-end of capped RNAs, and their variations are being used for a wide range of assays and contributing to unveil novel species and/or structures of RNAs.

Here I would like to introduce a project, FANTOM4 (Functional Annotation of Mammalian Genomes 4) using a next generation sequencer, where we adopted CAGE to profile promoters during a single differentiation time course of a human leukemic cell line, THP-1, and performed an analysis of transcriptional regulation. We complemented the analysis with other high-throughput experiments to provide further evidences of the profiled promoters and the predicted regulations. All the results are integrated into a single resource, the FANTOM web resource, where we included explicit and standardized description (metadata) of the experiments, such as protocols, samples, and data files of the identified entities (promoters) with quantification (expression). I also present some difficulties we faced in the process of integration and the solution we adopted, as a case study of elaborate experiment annotations. Further efforts based on this type of experimental evidences would be required to annotate all the complexity of transcriptome, ultimately.

UniRule - Automatic Annotation In UniProtKB

Alan Bridge1, Diego Poggioli2, Claire O'Donovan2, and the UniProt consortium1,2,3

1 Swiss Institute of Bioinformatics, Centre Medicale Universitaire, 1 rue Michel Servet, CH1211 Geneva 4, Switzerland
2 EMBL European Bioinformatics Institute, Wellcome Trust Genome Campus, Hinxton, Cambridge CB10 2SD, UK
3 Protein Information Resource, Georgetown University Medical Center, 3300 Whitehaven St. NW, Suite 1200, Washington, DC 20007, USA

The UniProt KnowledgeBase (UniProtKB) provides a stable, comprehensive, freely accessible, centralized resource on protein sequences and functional annotation. UniProtKB consists of two sections: UniProtKB/SwissProt, a section containing records that are manually annotated with information extracted from the literature and curator-evaluated computational analysis, and UniProtKB/TrEMBL, a section containing computationally analysed records enriched with automatic annotation and classification. Automatic annotation is an essential complement to manual annotation, which cannot keep pace with either current or projected rates of growth of UniProtKB. To this end UniProt is developing an integrated annotation system termed UniRule, which is based on both manually curated and automatically generated annotation rules and signatures. UniRule provides a variety of annotation types including protein names, general functional annotation in the form of free text and controlled vocabularies, sequence annotation including domains and residues of functional importance, and inferred family relationships. At the time of writing (UniProtKB release 2010_08), UniRule provides annotations for around 35% of more than 11 million UniProtKB/TrEMBL entries. The UniRule system and plans for its future development will be presented.
Community Annotation in Biology
Raja Mazumder
Protein Information Resource, Department of Biochemistry and Molecular & Cellular Biology, Georgetown University Medical Center, USA

Attempts to engage the scientific community to annotate biological data (such as protein/gene name and function) stored in databases have not been overly successful. There are several hypotheses on why this has not been successful but it is not clear which of these hypotheses are correct. We surveyed 50 biologists (who have recently published a paper characterizing a gene or protein) to better understand what would make them interested in providing input/contributions to biological databases. Based on our survey two things become clear: a) database managers need to proactively contact biologists to solicit contributions; and b) potential contributors need to be provided with an easy-to-use interface and clear instructions on what to annotate. Other factors such as 'reward' and 'employer/funding agency recognition' previously perceived as motivators was found to be less important. Based on this study and feedback from participants in bioinformatics courses and workshops, we propose that community annotation projects should devote resources to direct solicitation for input and to streamlining the processes and interfaces used to collect this input.

Nanopublication by Community Annotation in the ConceptWiki
Christine Chichester, Kees Burger, Rob Hooft, Barend Mons
Netherlands Bioinformatics Centre, Nijmegen, The Netherlands

Time and attention are finite. Biocuration is a necessary time and labor-intensive task, which is however unavoidable to make high quality biological information accessible to both humans and computers. To tackle these opposing paradigms, we have developed a system that can reduce biological information to the scale of ‘micro-blogs’. These snippets of biological information, in the form of richly annotated subject-predicate-object triples called nanopublications, can be simply and easily produced in the ConceptWiki (www.conceptwiki.org) and attributed to the creating author. On each Concept page in the ConceptWiki (e.g. for each protein, disease, biological process, etc.), any registered curator can select, from dropdown menus, suitable predicate-object combinations to create nanopublications. The newly constructed nanopublication will be recorded on the authors' individual page giving the appropriate intellectual credit to the creator. Additionally, other ConceptWiki registered authors previously indicating interest in the concepts present in a nanopublication will be alerted to its creation. This provides a channel to otherwise inaccessible resources and allows users to connect to each other, the social glue of community annotation. In addition, nanopublications that were created via text mining will be linked to the Subject and Object pages and can be candidates for community annotation as well. These methods for biocuration have been established in the ConceptWiki knowing that participation in community annotation is motivated by a number of factors: accolades, knowledge gain, fun, and fulfilment. Community annotation initiatives should offer participants a mix of all these factors to be successful.
The Challenge Of Eukaryote Genome Annotation At Genoscope

Betina M. Porcel, Franck Aniere, Sylvain Bonneval, Benjamin Noel, Jean-Marc Aury, Corinne Da Silva, Olivier Jaillon, France Denoeud, Claude Searpelli, Jean Weissenbach, Patrick Wincker and François Artiguenave.
Genoscope, Institut de Génomique, CEA, 2 Rue Gaston Crémieux, CP 5706, 91057 EVRY cedex, France

Annotating a genome aims at assigning valuable biological information on genomic sequences. The annotation process includes different steps, starting by the definition of the gene structures and ending by the functional annotation of protein functions and the definition of homology relationships. Finding these genes involves computational methods as well as experimental validation.

As a consequence of the adaptation of the novel generation sequencing machines, the number of eukaryote genomes sequenced and annotated at Genoscope is increasing constantly. For the last 5 years, we have been annotating over 10 eukaryote genomes using our Eukaryote Genome Annotation workflow. We will present few examples of annotated genomes as well as our developments to automate the annotation process which allow us to annotate today 2 to 4 genomes a year.

However, even though automatic computational predictors are extremely useful for a large scale annotation at a preliminary level, their predictions are sometimes inadequate. Therefore, any gene models require human expertise to find errors and resolve incongruous evidence on the automatic annotation of the genome. In order to achieve both high-throughput and high quality of annotations, we set up a scalable model for Community Genome Annotation, which combines automated annotation, community-wide genome analysis and manual validation based on GMOD components (Chado/GBrowse/Apollo). We provide to our collaborators a distributed annotation platform allowing an expert evaluation of genome annotations, in addition to our automated gene prediction pipeline.

Training biocurators in annotation jamborees, as well as coordination and support for the user communities have been launched, as it will be shown. This community based-manual annotation facilitates expert networking and comparison of predicted annotations with existing biological data.

The Neuroscience Information Framework: A Unified Semantic Framework for Discovery and Integration of Biomedical Data and Resources on the Web


Ctr. for Res. in Biol. Systems, UCSD, USA; Neurobio., Yale Univ., USA; Anat. and Neurobiolog, Washington Univ., St Louis, USA; Krasnow Inst. for Advanced Study, George Mason Univ., USA; Caltech, USA

Informatics and new web technologies are becoming increasingly important to biomedical researchers. An initiative of the NIH Blueprint for Neuroscience Research, the Neuroscience Information Framework (NIF; http://neuinfo.org) enables discovery and access to public research data, contained in databases and structured web resources (e.g. queryable web services) that are sometimes referred to as the deep or hidden web, and resources through an open source dynamic inventory of biomedical resources that are annotated and integrated with a unified system of biomedical terminology. The NIF Database Federation (with more than 60 independent databases) allows for direct search, discovery and integration of database content. Search and annotation of resources and resource content is enhanced through the utilization of a comprehensive ontology (NIFSTD; http://purl.org/nif/ontology/nif.owl), built as a set of modular ontologies. To enable broad community contribution to NIFSTD, NeuroLex (http://neurolex.org) is available as a wiki that provides an easy entry point for the community. New services being provided include a complete full text index of PubMed Central’s Open Access articles, an annotation framework (allowing content from the NIF data federation to be annotated by the NIFSTDTOntologies) that allows data resources to be efficiently indexed (Lucene) for user searches and also made available as a SPARQL (RDF) end-point, entity highlighting services, and informational pop-ups (NIF Cards) that can be linked into any application. NIF cards draw upon NIFSTD and the NIF data federation to display information about an entity and provide customized search options depending upon the domain. As the NIF Cards evolve, they will provide the basis for linking results into the large ecosystem of linked data. In addition, enhancements have been made to the NIF ontologies, including the ability to intelligently handle inferred classes such as “GABAergic neuron”.

28
Incorporating Community Annotation Interfaces into the CIPRO2.5 Database with Comprehensible Sketches to Support Quick Annotations of Proteome Data.

Keisuke Ueno¹, Kouki Yonezawa¹, Katsuhiko Mineta¹, Toshinori Endo¹

¹Graduate School of Information Science and Technology, Hokkaido University, North 14 West 9, Sapporo, Hokkaido 060-0814, Japan

User annotation or comment interfaces are now widely used in several web sites such as journals, news, weblogs and Wikipedia. However, there are only a few biological databases with annotation interfaces. The Ciona intestinalis protein database (CIPRO) was created in order to provide integrated proteome data especially for experimental biologists. The current database contains 89,673 unique sequences covering all the known and predicted gene models. Typical tasks include which gene models are reliable and which function is plausible. The human-curated annotation is most important for the meaningful database.

Here we incorporate three new functions into the CIPRO2.5 database (http://cipro.ibio.jp/2.5/), providing enriched resources for the users. First, a community annotation interface as web forms and a user comment editor with rating its comment were added. Second, the web pages were specifically designed to compact for quickly understood overviews. For example, cytolocalization was automatically provided by a color-depicted cell image based on the intensity instead of the numeric values of raw data. The expression data of EST, microarray and 2D-PAGE were integrated as one chart. In addition to these data, the images of transmembrane prediction, domain and motif search, and the OMIM ortholog on the chromosome map were included in each protein page. The last, even the BLAST and PMF search were added to combined fields in the retrieval system.

As a result, a total of 11,134 pages were annotated by our community. Furthermore, 2,186 comments were added to the database. Those annotated data are freely accessible at the CIPRO2.5 web site.
Combining Computational Prediction and Manual Curation to Create Plant Metabolic Pathway Databases

Peifen Zhang, Kate Dreher, Anjo Chi, Lee Chae, Cynthia Lee, and Seung Yon Rhee

Department of Plant Biology, Carnegie Institution, Stanford, California 94305, USA

Reconstructing metabolic networks from genome sequences enables systems-level analyses of genomes in a metabolic context. The goal of our project, entitled the Plant Metabolic Network (PMN, plantcyc.org), is to develop a general approach for computationally reconstructing high quality metabolic pathway databases for plants.

The PMN infrastructure includes a sequence annotation pipeline for predicting enzyme functions of protein sequences, a computational prediction of species-specific metabolic networks from annotated enzymes using the Pathway Tools software, and an automated pathway validation procedure. To tailor the platform for plant genomes, we have developed an extensively curated plant-specific biochemical pathway database called PlantCyc that is used as a reference by the Pathway Tools software. PlantCyc catalogs a comprehensive set of known plant biochemical pathways involved in primary and secondary metabolism. The first version of the PMN infrastructure has been applied to the recently sequenced genomes of poplar and soybean resulting in the creation of PoplarCyc and SoyCyc.

Major types of information curated in PlantCyc, and other PMN databases, include pathway diagrams and summaries, reaction equations, compound chemical structures, and enzyme physicochemical properties. Evidence codes along with references are attached to pathways and enzymatic reactions for quality assurance. In the near future, we will improve the sequence annotation pipeline and expand the types of curated information to include rate-limiting steps of pathways and transcriptional regulators of enzymes.

KaPPA-View4: A Pathway Database for Gene Co-expression and Metabolite Co-accumulation Analysis.

Nozomu Sakurai1, Daisuke Shibata1

1Kazusa DNA Research Institute, Kisarazu, Chiba, Japan

Correlations of gene-to-gene co-expression and metabolite-to-metabolite co-accumulation are considered to be useful for uncovering unknown functions of genes and regulatory systems of the metabolic pathways. Although many databases and tools are available to interpret quantitative data of transcriptome and metabolome, there are only limited ones that connect correlation data to biological knowledge and can be utilized to find biological significance of it. We developed a new metabolic pathway database, KaPPA-View4 (http://kpv.kazusa.or.jp/kpv4/), that enables to overlay gene-to-gene and/or metabolite-to-metabolite relationships as quadratic Bezier curves on metabolic pathway maps. For up to 4 maps including pathway maps, gene category maps, and user created maps can be analyzed on a single browser window, the representation would help to discover, for example, novel functions of a transcription factor that regulates genes on a metabolic pathway. Pathway maps of the Kyoto Encyclopedia of Genes and Genomes (KEGG) and maps generated from their gene classifications are available at KaPPA-View4 KEGG (http://kpv.kazusa.or.jp/kpv4-kegg/). One of the major advantages of utilizing KEGG pathway maps is that we can share the latest results of KEGG’s continuous effort for curation of gene descriptions, categorizations, and assignment on the maps for various organisms. We currently provide pathway map data for 20 species including animals, plants and microorganisms, and gene co-expression data for 12 species that are retrieved from ATTED-II, COXPRESdb, CoP and MiBase.

This work was supported by the New Energy and Industrial Technology Development Organization (NEDO, Japan).
MassBank: A Reference Database for Chemical Identification of Metabolites Detected by Mass Spectrometry

Takaaki Nishioka\textsuperscript{1,2}

\textsuperscript{1}Institute for Advanced Biosciences, Keio University, Tsuruoka, Yamagata 997-0035, Japan
\textsuperscript{2}BIRD, Japan Science and Technology Agency, Tokyo 102-0081, Japan

MassBank (http://www.massbank.jp/) is the first public repository of mass spectral data of biological, and chemically synthetic small molecules. Contributors prepare their spectral data in a common record format, “MassBank record format” and publish the formatted data from their own data servers. ESI-MS\textsuperscript{2} spectra contributed to MassBank were analyzed by 14 different, well-controlled analytical methods. However, slight differences in their experimental methods gave different mass spectra to identical chemical compounds, which are not suitable as the reference data for the identification of metabolites detected by mass spectrometry. To solve this problem, MassBank provides one artificially merged ESI-MS\textsuperscript{2} spectrum for each chemical compound. They were prepared by overlaying and merging the ESI-MS\textsuperscript{2} spectra of identical chemical compounds. As their yielding representative fragmentation patterns, merged ESI-MS\textsuperscript{2} spectra as the reference data significantly improved the precision in the identification by using a MassBank service, “Spectral Similarity Search”. Another service specific to MassBank, “Peak Search Advanced”, is to retrieve merged ESI-MS\textsuperscript{2} spectra that gave product ions or neutral losses specified by molecular formulae. We manually annotated molecular formulae to more than 4,000 major product ions and predicted the chemical substructures that product ions were derived from. By using the service, for example, we can retrieve the spectra that give a neutral loss of C\textsubscript{6}H\textsubscript{12}O\textsubscript{6} (180.0633 Da) but not of C\textsubscript{7}H\textsubscript{8}N\textsubscript{4}O (180.0647 Da). The two neutral losses, which are different only by 0.0014 Da (8 ppm), are difficult to identify by conventional search methods. In combination with “Substructure Search”, “Peak Search Advanced” is useful to analyze the chemical relations between product ions and chemical substructures. Accumulations of such relations are essential to elucidate the chemical structure of unidentified chemical compounds detected by mass spectrometry.

Plant Metabolic Data Curation And Its Integration Into Solanaceae Genomic Databases

Anuradha Pujar\textsuperscript{1}, Ron Caspi\textsuperscript{2}, Naama Menda\textsuperscript{1}, Isaak Tecle\textsuperscript{1}, Aureliano Gomez\textsuperscript{1}, Peter Karp\textsuperscript{2}, Lukas Mueller\textsuperscript{1}

\textsuperscript{1}Boyce Thompson Institute, New York, USA
\textsuperscript{2}Stanford Research Institute, California, USA

Curation of plant metabolic data presents several challenges and the MetaCyc database has developed a sophisticated curation software suite called Pathway Tools software. This software enables the curation of complex metabolic information starting from metabolic compounds, reactions, enzymes, genes and entire pathways. The data visualization is interactive and allows the viewing at several levels of metabolism, from structural details of a single plant compound to superpathways spread across several plant species. MetaCyc is the encyclopedia of metabolic data of all organisms, linkages to genomic information is done via the Pathway Genome Databases (PGDB’s). Here we describe the integration of Solanaceae metabolic data with the genome database. The metabolic information of the Solanaceae family is rapidly increasing and large numbers of pathways are being characterized, simultaneously high throughput studies on their metabolome and transcriptome are also being done. The availability of the Tomato genome sequence provides leverage to connect disparate metabolic information to the gene sequence, genetic resources and phenotypes. At the Solanaceae Genomics Network (SGN) (http://www.solgenomics.net), an interactive platform was created that interfaces metabolic data with genomic and other kinds of information, MetaCyc, and annotated data from SGN. SolCyCG (http://www.solgenomics.net/tools/solcyc/) is a collection of Pathway Genome Databases, developed for the clade oriented Solanaceae Genomics Network database. It has predicted metabolic pathway databases of significant Solanaceae species and includes Lycocyc (tomato), SolaCyc (eggplant), NicotianaCyc (tobacco), PetuniaCyc (Petunia), CucCyc (Capsicum) and PotatoCyc (potato). Newly curated metabolic pathways of Solanaceae include; flavonoid and wax metabolism of Tomato fruit surfaces, tropane alkaloid pathways of Solanum species, GABA shunt metabolism during fruit ripening, tomato steroidal glycoalkaloid pathway and secondary metabolite biosynthesis in Solanaceae trichomes.
The EcoCyc Database – Integrating And Transferring Knowledge About E. coli

I.M. Keseler¹, T. Altman¹, C. Bonavides-Martinez², J. Collado-Vides², C. Fulcher¹, S. Gama-Castro², R. Gunsalus³, A. Kothari¹, M. Krummenacker¹, M. Latendresse¹, A. Mackie⁴, S. Paley¹, I. Paulsen⁴, M. Peralta-Gil², A. Santos-Zavaleta², M. Sarker¹, A.G. Shearer¹, V. Weiss², and P.D. Karp¹

¹SRI International, Menlo Park, CA, USA, ²Center for Genomic Sciences, Natl. Autonomous Univ. of Mexico, Cuernavaca, Mexico, ³Univ. of California, Los Angeles, CA, USA, ⁴Macquarie University, Sydney, Australia

EcoCyc (EcoCyc.org) is a model organism database for Escherichia coli K-12 MG1655. Since 1994, expert curators have captured information found in the experimental literature, resulting in extensive coverage of the functions of individual E. coli gene products. The Pathway Tools software provides both the curation interface to curators, as well as search and display tools for access by end users.

From its roots as a metabolic pathway database, EcoCyc has continuously expanded its coverage to include many more data types, which is enabled by a corresponding expansion of the database schema and Pathway Tools software. These enhancements have been driven by the availability of new types of experimental data, requests from the user community, and internal development of novel bioinformatics tools. Major recent enhancements include tools for curation and representation of regulatory mechanisms that act after transcription initiation, such as regulation by attenuation, riboswitches, and regulation by small RNAs. A new interactive editor has greatly improved the ability to curate and display signaling pathways.

We have expanded our collection of databases built from the genome sequences of other strains of E. coli and Shigella. Within the BioCyc collection at BioCyc.org, more than 30 E. coli and Shigella genomes are currently represented, allowing comparative analyses between different pathogenic, non-pathogenic and laboratory strains of E. coli. However, as increasing numbers of microbial genome sequences are becoming available, the problem of inadequate, inconsistent or outdated annotation is becoming more evident. A future enhancement to the Pathway Tools software will include tools for curators to transfer annotations from a well-curated model-organism genome to the less-well curated relatives, including pathogenic organisms.
Curation for Systems Biology Resources in an Academic Model

Keshava Prasad T. S.

Institute of Bioinformatics, International Tech Park, Bangalore 560 066, INDIA

Systems biology approaches are increasingly utilized in biological research. Molecular genetic investigations have transformed from being small scale studies targeted at a single gene/protein to high-throughput discovery platforms. The advantage of these discovery platforms is that they provide an unbiased means to identify genes/proteins associated with various biological processes and/or diseases. However, to efficiently harness the power of systems biology, we need to establish strong ‘building blocks’ of the systems. A set of well-annotated integrated data resources and validated genomic and proteomic datasets can together represent such building blocks. Although computational approaches have been utilized in text mining and data acquisition from published literature, interpretation by trained biologists are necessary to competently curate plethora of heterogeneous biological information. At the Institute of Bioinformatics (http://www.ibioinformatics.org) in Bangalore, we experimented with an academic model where we have involved Ph.D. level scientists and doctoral students in curating biological data. We have developed a number of databases and utilities to curate the data. These include Human Protein Reference Database, Human Proteinpedia – a proteome resource with distributed annotation system, NetPath- a database of signaling pathways and PhosphoMotif Finder. Biologists who are trained in curation have effectively utilized the knowledge gained to design and investigate gene and protein expression profiling in human diseases. The databases developed by us are now being enriched with the vast amount of high-resolution data generated in our laboratory, which will further aid systems level discoveries.

Identification Of Causative Variations Based On The Technology Of Next Generation Sequencing

Yixue Li

1 Shanghai Center For Bioinformation Technology Qinzhou Rd. 100, 200235 Shanghai, China

Next generation sequencing is driving growth within the basic and biomedical research communities as rapidly as the bases are being sequenced. Many researches in life sciences can be accelerated dramatically by this kind of deep sequencing technologies. Now Chip-Seq, Methyl-Seq, RNA-Seq, whole exome capture etc. have already been as popular as other technologies like biochips, and probably will replace the latter soon. All agree, however, that with the implementation of any new technology there is a balancing act of cost-quality-quantity. One of potential applications of the next generation sequencing is whole exome capture, if we do want to balance act of cost-quality-quantity. This deep sequencing based technology offers a good opportunity to help us decipher disease related variations by smaller samples and lower costs comparing with that of GWAS study. Currently, newly launched Solexa and Solid sequencers can produce 200-300GB bases per run with much lower price. Based on the new approaches many research laboratories try to use whole genome sequencing for few individuals to identify causative variations of diseases if he can prepare good samples and perform better bioinformatics studies and have large cohorts for the results evaluation.
ASIAN MODELS OF MICROBIAL RESOURCE INFORMATION NETWORK

Juncai Ma

Information Network Center, Institute of Microbiology, Chinese Academy of Sciences, China No. 1 West Beichen Road, Chaoyang District, Beijing 100101, P. R. China. ma@im.ac.cn

Abstract

The report introduces Asian models of microbial resource information network, including the development of ABRCN, RNAM and BOLD system. Based on the microbial resources of China General Microbiological Culture Collection Center (CGMCC), Institute of Microbiology, Chinese Academy of Sciences (IMCAS) established the China Information Network of Biological Resource Center (BRC). In 2005, CGMCC of China, NBRC, KCTC and BIOTEC completed Asian Biological Resource Center Network (ABRCN), and a data portal for public was set up. In 2010, Chinese Academy of Sciences (CAS) launches Research Network for Applied Microbiology (RNAM). One of its tasks is setting up Information Center for RNAM, which will develop virtual lab and network platform for joint research, and establish the project management platform of RNAM. It will also develop CAS Information Portal of RNAM. Barcode of Life is an international collaborative program aiming at collecting and identifying as much as possible creatures via relatively short DNA segments (DNA barcodes). As a central node, China is setting up the data system that will be an important complement to BOLD systems in Canada.

Biocuration of Genomic Data of Non-model Species in Taiwan

Chao-Li Huang¹, Chuan-Wen Ho¹, Chi-Chun Huang¹, Wei-Kuang Wang¹, Tsai-Wen Hsu², Tzen-Yuh Chiang¹

¹Department of Life Sciences, National Cheng-Kung University, Tainan, Taiwan 701
²Taiwan Endemic Species Research Institute, Nantou, Taiwan 552

Next-generation sequencing technologies enable one to look at genomic divergence between species. De-novo sequencing of both genomic DNA and transcriptome nevertheless creates difficulties and challenges in biocuration of sequence data. Here, we present genomic analyses of wild Arabidopsis halleri ssp. gemmifera that is affined to model species of A. thaliana and A. lyrata, and Miscanthus, a biofuel plant phylogenetically distant to model species in the Poaceae. Genomic DNA of A. halleri ssp. gemmifera was sequenced with a 454 sequencer. In total, 21,361 contigs were assembled, with 13,055 (61.1%) related to homologous genes with annotated functions, and 4,288 (20.1%) were identified as intergenic spacers, while another 4,018 (18.8%) remained unidentified. We randomly selected 98 genes for examining the genetic divergence between species. Coalescent analyses suggested that the segregation of ancestral polymorphisms alone cannot explain the high inconsistency between gene trees across loci. Rampant gene introgression distorts the molecular phylogenies of Arabidopsis species. However, not all genes migrated across species freely. Gene ontology analyses suggested that some non-migrating genes were constrained by natural selection. A transcriptome of a nonmodel species of Miscanthus sinensis was constructed with over 7G bases obtained from Solexa and 454 sequencing. Analyzing 6,491 contigs after Solexa and 454 reads assemblage, we've found 2,854 distinct blast best hits. Comparing with the model species of sorghum and rice, about homologs of 86.7% and 80.6% contigs can be identified. Of them, 592 contigs were identified as transcription factors and 109 contigs are related to stress-response genes. In contrast to most nuclear genes that were shaped by negative selection, Ka/Ks >1 detected at some other genes like Heat-Shock Protein 70 locus and genetic differentiation between populations suggested a critical role in local adaptation.
The authors introduce the database and Graphical User Interfaces (GUIs) of Genome Network Platform and its application to comprehensive analysis on transcription network. Genome Network Platform has been developed by National Institute of Genetics as a research infrastructure for Genome Network Project (GNP), which is headed by Ministry of Education, Culture, Science and Technology. This platform is designed to provide integrated data that helps elucidating biological networks including transcriptional regulation. There are accumulating data produced by GNP consortium members; genomic information such as transcription start sites and transcription regulatory region, transcriptomic information such as expression and splice variants, proteomic information such as domain structures and protein-protein interactions. Genome Network Platform contains these data produced by GNP consortium members and public data. It will be useful to integrate these data successfully for the researchers in areas including a comprehensive analysis of biological process, but we have some difficulties to integrate and utilize simply the different type of data such as genome, transcriptome and proteome. Moreover the amount of data is too huge to handle with ease, something of summarize is required to handle in all. The authors propose the data integration model and GUIs that are focus on the gene transcription. Here, we explain the model and demonstrate the values of GNP GUIs for study on transcription network.
Omics Data Sharing

Susanna-Assunta Sansone¹ and Dawn Field²

¹University of Oxford Oxford e-Research Centre, UK;
²Natural Environment Research Council, Environmental Bioinformatics Centre, UK

Several data preservation, management and sharing policies have emerged in response to increased funding for high-throughput approaches in major genomics and functional genomics bioscience domain; and nowadays, several funding agencies also require inclusion of data-sharing plans in grant applications. But despite some commonalities, the policies are heterogeneous by nature, given the different types of communities served and the data types they generate. In parallel, an escalating number of community-driven standardization efforts (including biocurators, database developers and experimentalists, vendors etc), operate to develop minimal requirements checklists, ontologies, and file-formats to support the harmonization of the reporting process, so that different experiments and data can be easily shared, compared, and integrated. The proliferation of these standardization efforts is a positive sign of community engagement, but it also brings with it new sociological and technological challenges - creating interoperability and avoiding unnecessary overlaps and duplication of efforts that hampers their wider uptake. Let aside the ethical, commercialization, credit and other known issues arising from public data release, basic communication channels still need to be formally created and maintained, especially between the funders and the standardization efforts to enable flow of information and mutual support. For example, funding agencies should keep abreast of challenges the standardization efforts face so that they can provide targeted funds to sustain their development and maintenance; when standards-compliant systems become available these should be channeled to the appropriate funding agencies that in turn can recommend them in the data sharing policy. Hence the creation of the BioSharing forum on data policies’s plans and reporting standards, stemming from an initial collaboration with a range of representatives from US, UK and European funding agencies (Field D., Sansone SA, et al. (2009), Science, 9, 234-236; www.biosharing.org); our dialogue will continue to formalize linkages between funders, implementing data sharing policies, and well-constituted standardization efforts in the biosciences domain and in close collaboration with the Biocuration Society.

H-InvDB: A Comprehensive Annotation Resource For Human Transcriptome

Chisato Yamasaki¹, Jun-ichi Takeda¹, Takuya Habara¹, Makoto Ogawa²,³, Akiko Noda¹, Ryuichi Sakate¹, Katsuhiko Murakami², Tadashi Imanishi¹ and Takashi Gojobori¹,⁴

¹Biomedicinal Information Research Center, National Institute of Advanced Industrial Science and Technology, Japan
²Japan Biological Informatics Consortium, Japan
³DYNACOM Co., Ltd., Japan
⁴National Institute of Genetics, Japan

H-Invitational Database (H-InvDB: http://www.h-invitational.jp/) is a comprehensive annotation resource for human transcriptome. By extensive analyses of all human transcripts, we provide curated annotations of human genes, transcripts and proteins that include gene structures, alternative splicing isoforms, non-coding functional RNAs, protein functions, functional domains, sub-cellular localizations, metabolic pathways, protein 3D structure, genetic polymorphisms, relation with diseases, gene expression profiling, molecular evolutionary features, protein-protein interactions (PPIs) and gene families/groups. The latest release of H-InvDB (release 7.0) provides annotation for 296,912 human transcripts in 46,499 human gene clusters based on human full-length cDNAs, mRNAs and the reference human genome sequences (NCBI b37.1). H-InvDB consists of three main views, the Transcript view, the Locus view and the Protein view, and six sub-databases; G-integra, H-ANGEL, DiseaseInfo Viewer, Evola, PPI view and Gene Family/Group view. We also provide data mining tools such as “Navigation search”, an extended search system that enables complicated searches by combining 16 different search options (http://www.h-invitational.jp/hinv/c-search/hinvNaviTop.jsp) and “H-InvDB Enrichment Analysis Tool (HEAT)”, a data mining tool for automatically identifying features specific to a given human gene set (http://hinv.jp/HEAT/).
Literature Curation of Protein Interactions: Measuring Agreement Across Major Public Databases

Andrei L. Turinsky\(^1\), Sabry Razick\(^2,3\), Brian Turner\(^1\), Ian Donaldson\(^2,4\) and Shoshana J. Wodak\(^1,5\)

\(^1\) Molecular Structure and Function program, Hospital for Sick Children, Toronto, Canada
\(^2\) The Biotechnology Centre of Oslo, University of Oslo, Norway
\(^3\) Biomedical Research Group, Department of Informatics, University of Oslo, Norway
\(^4\) Department of Molecular Biosciences, University of Oslo, Norway
\(^5\) Department of Molecular Genetics, Department of Biochemistry, University of Toronto, Canada

Literature curation of protein interaction data faces a number of challenges. Although curators increasingly adhere to standard data representations, the data that various databases actually record from the same published information may differ significantly. Some of the reasons underlying these differences are well known, but their global impact on the interactions collectively curated by major public databases, has not been evaluated.

Here we quantify the agreement between curated interactions from 15,471 publications shared across nine major public databases. Results show that on average, two databases fully agree on 42% of the interactions and 62% of the proteins curated from the same publication. Furthermore, a sizable fraction of the measured differences can be attributed to divergent assignments of organism or splice isoforms, and to alternative representations of multi-protein complexes. Our findings highlight the impact of divergent curation policies across databases, and should be relevant to both curators and data consumers interested in analyzing protein-interaction data generated by the scientific community. Availability: [http://wodaklab.org/iRefWeb](http://wodaklab.org/iRefWeb)

neXtProt, a New Human-Centric Protein Knowledge Resource

Lydie Lane

University of Geneva and Swiss Institute of Bioinformatics, Switzerland

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 ~20'300 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 from the University of Geneva and the SIB, we are developing neXtProt, a human-centric protein knowledge resource. We aim to help researchers answer pertinent questions relevant to human proteins. This requires answering three 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 highthroughput approaches such as proteomics, microarray, antibodies, 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. Finally, we also aim in enabling researchers to collaborate and share data and ideas through the use of the neXtProt platform. neXtProt is a common development of the SIB and of GeneBio SA. We hope you will enjoy working with it and will help us making it evolve by telling us of your specific needs.
UniProt Knowledgebase: A Hub Of Integrated Protein Data

Michele Magrane1, UniProt Consortium1,2,3

1EMBL-European Bioinformatics Institute, UK
2Swiss Institute of Bioinformatics, Switzerland
3Protein Information Resource, USA

Data integration plays an increasingly important role in bringing together the large amounts of diverse information spread across disparate resources and presenting a comprehensive overview of these data to the scientific community. The UniProt Knowledgebase (UniProtKB) acts as a central hub of protein knowledge by providing a unified view of protein sequence and functional information. Manual and automatic annotation procedures are used to add data directly to the database while extensive cross-referencing to more than 120 external databases provides access to additional relevant information in more specialised data collections. UniProtKB also integrates data such as protein sequences, protein-protein interactions, Gene Ontology terms and official gene nomenclature from a range of resources. All information in UniProtKB is attributed to its original source, allowing users to trace the provenance of all data. In addition, UniProtKB data is made freely available in a range of formats to facilitate integration with other databases and the UniProt Consortium is committed to using and promoting common data exchange formats and technologies. This approach ensures that information is captured in the most appropriate resource for subsequent integration with other databases and also ensures maximum curation efficiency by preventing duplication of efforts across multiple resources. How UniProt achieves this data capture and integration will be presented. The UniProt resource is available at www.uniprot.org.
MetaCuration Standards and Minimum Information about a Bioinformatics Investigation

Tin Wee TAN¹,², AM Khan¹,², SJ Lim¹, YL Hu¹, M de Silva¹, KS Lim¹
¹National University of Singapore, Singapore, ²Asia Pacific Bioinformatics Network

Many bioinformatics databases published in journals are here this year and gone the next. There is generally (i) no requirement, mandatory or otherwise, by reviewers, editors or publishers for full disclosure of how databases are built and how they are maintained; (ii) no standardized requirement for data in public access databases to be kept as backup for release and access when a project ends, when funds expire and website terminates; (iii) the case of proprietary resources, there is no requirement for data to be kept in escrow for released under stated conditions such as when a published database disappears due to company closure. Consequently, much of the biological databases published in the past twenty years are easily lost, even though the publications describing or referencing these databases and webservice remain. Given the volume of publications today, even though it is practically possible for reviewers to re-create databases as described in a manuscript, there is usually insufficient disclosure and raw data for this to be done, even if there is sufficient time and resources available to perform this. Consequently, verification and validation is assumed, and claims of the paper accepted as true and correct at face value. A solution to this growing problem is to experiment with some kind of minimum standards of reporting such as the Minimum Information About a Bioinformatics Investigation (MIABi) and standardized requirements of data deposition and escrow for enabling persistence and reproducibility. With easy availability of cloud computing, such a level of reproducibility can become a reality in the near term. Through standards in meta-curation and minimum standards of reporting that uphold the tenets of scientific reproducibility, verifiability, sustainability and continuity of data resources, the knowledge preserved will underpin tomorrow's scientific research. Other issues include disambiguation of authors or database names, and unique identifiers to support non-repudiability, possibly in multiple languages. The International Conference on Bioinformatics and its publications are now in the process of making attempts at addressing these issues and this presentation will highlight some of the current efforts.

Utopia Documents : Linking Literature And Data

Steve Pettifer, Dave Thorne, Teresa Attwood, Philip McDermott, James Marsh

The University of Manchester, United Kingdom

In recent years, the gulf between the mass of accumulating research data and the massive literature describing and analysing those data has widened. The need for intelligent tools to bridge this gap, to rescue the knowledge being systematically isolated in literature and data silos, is now widely acknowledged.

To this end, we have developed Utopia Documents, a novel PDF reader that semantically integrates visualisation and data-analysis tools with published research articles. In a successful pilot with editors of the Biochemical Journal (BJ), the system has been used to transform static document features into objects that can be linked, annotated, visualised and analysed interactively (http://www.biochemj.org/bj/424/3/). Utopia Documents is now used routinely by BJ editors to annotate and enhance article content prior to publication. Recent additions include integration of various text-mining and biodatabase plugins, demonstrating the system’s ability to seamlessly integrate on-line content with PDF articles.
AMIS, The Article Minimum Information Standard

Delphine Dauga¹, Patrick Lemaire¹

¹ Institut de Biologie du Développement de Marseille-Luminy, UMR6216 CNRS, F-13288 Marseille Cedex 9, France

The curation process is significantly slowed down by missing information in the articles analyzed (for example, the identity of the clones used to generate ISH probes, the precise sequences tested in reporter assays, etc...). To help authors ensure in the future that necessary information is present in their article, we defined the Article Minimum Information Standard (AMIS) guidelines. This standard describes for each experiment the mandatory information that should be mentioned in literature articles to facilitate the curation process. These guidelines extend the minimal information defined by the MISFISHIE format (Deutsch et al. 2008, Nature Biotechnology). This standard was deduced from the ANISEED curation pipeline (Tassy, Dauga, Daian, Sobral et al. 2010, Genome Research). ANISEED is a generic infrastructure for the creation, maintenance and integration of molecular and anatomical information on ascidians. Thanks to the ANISEED curation pipeline, the capture of published information was streamlined by the creation of the “Article Card” concept. Each Article Card summarizes in a standardized and structured format the content of the text and figures of an article. It lists, and links to the corresponding experimental evidences, all features studied (genes, cell fates, etc...). This curation strategy allowed pointing out missing information essential to transform the “biological interpretability” of the data into their “computability”. AMIS was defined to obviate this problem. The MISFISHIE format doesn’t include the minimal information necessary to describe cis-regulatory elements. In ANISEED, a sophisticated representation of the structure of cis-regulatory elements and their upstream regulators was designed. AMIS details the minimal information to describe a regulatory region. To facilitate regulatory region data transfer between databases, a document type definition (DTD) was developed, following the AMIS rules.

Publishing Interactive Articles: Integrating Journals And Biological Databases

Arun Rangarajan¹, Tim Schedl², Karen Yook¹, Juancarlos Chan¹, Stephen Haenel³, Lolly Otis¹, Sharon Faelt¹, Tracey DePellegrin-Connelly², Ruth Isaacson², Marek S. Skrzypek³, J. Michael Cherry³, Paul W. Sternberg¹,⁶ and Hans-Michael Müller¹

¹California Institute of Technology, USA, ²Washington University School of Medicine, USA, ³Dartmouth Journal Services, USA, ⁴Genetics Society of America, USA, ⁵Stanford University School of Medicine, USA, ⁶Howard Hughes Medical Institute, USA

In collaboration with the journal GENETICS, we've developed and launched a pipeline by which interactive full-text HTML/PDF journal articles are published with named entities linked to corresponding resource pages in WormBase (WB). Our interactive articles allow a reader to click on over ten different data type objects (gene, protein, transgene, etc.) and be directed to the relevant webpage. This seamless connection from the article to summaries of data types promotes a deeper level of understanding for the naïve reader, and incisive evaluation for the sophisticated reader. Further, this collaboration allows us to identify and collect information before the publication of the article. The pipeline uses automated recognition scripts to identify entities that already exist in the database and a self-reporting form we created at WB that is sent to the author by GENETICS for submitting entities that do not already exist in our database. We include a manual quality control step to make sure ambiguous links are corrected, and that all new entities have been reported and linked properly. The automated entity recognition scripts allows us to potentially link any object found in a database as well as to expand this pipeline to other databases. We have already adapted this pipeline for linking Saccharomyces cerevisiae GENETICS articles to the Saccharomyces Genome Database (SGD) and are currently expanding this pipeline for linking genes in Drosophila articles to FlyBase. By integrating journals and databases, we are integrating the major modes of communication in the biological sciences, which will undoubtedly increase the pace of discovery.
The SciKnowMine System: Infrastructure for Biocuration Applied To MGI Document Triage


USC / ISI, Marina del Rey, USA; University of Colorado, Denver, USA; The Jackson Laboratory, Bar Harbor, USA; University of Utah, Salt Lake City, USA

The development of tools to support literature-based biocuration based on text mining has generated only a few, notable examples of practical working systems (such as Textpresso) and these systems do not reflect cutting-edge technology from the Biomedical Natural Language Processing (BioNLP) community. We present SciKnowMine, an infrastructure project that is based around a centralized collaborative architecture to bring biocurators and NLP specialists together. We use Knowledge Engineering approaches to precisely model the biocuration workflow of the Mouse Genome Informatics (MGI) system. This provides important design criteria for our system. As a proof of concept, we are developing a prototype web-based interface that addresses the crucial biocuration task of ‘document triage’ (‘Is this paper of interest?’) in the context of the MGI workflow. We construct these tools within a component-oriented architecture (based on the Apache UIMA framework, the Unstructured Information Management Architecture; this is fast becoming a standard framework for sharing text-mining components, for instance within the U-Compare system). We are investigating mechanisms of parallelizing UIMA workflows to scale up processing to large numbers (millions) of documents and are constructing a set of standard library-, document- and phrase-based tools that may be disseminated via the web to biocurators at MGI. We will also conduct extrinsic evaluations to examine the effect of the tool’s usage within MGI to speed up curation in order to understand more deeply how high-performing NLP algorithms may best be used in the service of biocuration tasks. SciKnowMine is intended to provide a framework for text-mining support of biocuration that (a) scales effectively, (b) is closely tailored to the specific needs of biocurators, (c) can utilize a wide variety of NLP components, and (d) encourages NLP specialists to participate in BioNLP research of direct relevance to biocuration.
Comparative Protein Structure Modeling: An Effective Means To Explore Protein Function

Česlovas Venclovas
Institute of Biotechnology, Graičiūno 8, Vilnius, Lithuania

Three-dimensional (3D) structure is the most conserved feature of evolutionary related proteins. This observation is exploited in comparative modeling, in which experimental structures of related proteins are used as templates to construct 3D models of proteins of interest. Since function of globular proteins is determined by the 3D structure, protein models are a natural choice for exploring their function in the absence of an experimental structure. In addition to whatever can be obtained from sequence data, structural models, depending on their accuracy, may provide qualitatively different information such as geometrical features (shape, putative binding sites) or surface properties (charge distribution, hydrophobic patches).

The applicability of comparative modeling depends on the ability to detect related structures to be used as modeling templates. Once a suitable template is identified, the mapping of protein sequence onto the template structure becomes the major determinant of the model quality. We have been interested in both of these modeling issues for quite some time. Recently, we have developed a new distant homology detection method based on the profile-profile comparison. Tests showed that the new method is capable of detecting very remote similarities, in few cases outperforming structure-based methods. We have also been aiming at improving the second step in comparative modeling, the sequence-structure mapping, which has been identified during community-wide protein structure prediction experiments (CASP) as one of the bottlenecks. Our strategy to improve sequence structure alignments has relied on the identification of reliable regions of the alignment based on their stability followed by the refinement of unreliable regions using model assessment techniques. The methodology turned out to be one of the most successful during the last CASP. The main elements in this methodology will be discussed in more detail. The use of comparative models at different levels of detail (from structural domains to individual residues) to predict various functional properties will be illustrated with several biological examples.

Biomacromolecular Structure Annotation

Jasmine Young1, Atsushi Nakagawa2, Jawahar Swaminathan3, Hongyang Yao4, and wwPDB annotation team

1RCSB PDB, Rutgers University, New Jersey, USA, 2PDBj, Institute for Protein Research, Osaka University, Osaka, Japan, 3PDBe, European Bioinformatics Institute, Hinxton, UK, 4BMRB, BioMagResBank, University of Wisconsin Madison, Wisconsin, USA

The Protein Data Bank (PDB) is the single international repository for information about biological macromolecules whose structures have been determined by X-ray and neutron diffraction, NMR, and electron microscopy. The PDB currently contains more than 65,000 entries. The data deposited by authors contain the coordinates of the model and related experimental data, with information about source, biological sequence, and structure determination statistics. In order to provide uniform and high quality data, annotators process these data following standard procedures and guidelines established by the wwPDB international consortium of data collection and deposition centers (wwpdb.org). The annotation process includes validation checks of the chemistry and geometry of the polymer and small molecules, and checks of the fit of the model against experimental data. The validation reports created during annotation highlight any critical issues and suggest corrections to enable authors to fix any errors in the entry prior to release. Authors can also be easily share these reports with journal editors and referees to more fully evaluate the quality of a structure described in an article. The reports provide an assessment of a structure while keeping the coordinate file confidential, thereby protecting author proprietary concerns.

Based on many years of experience, the wwPDB is developing a common set of new deposition and annotation tools. These tools will adopt new validation procedures recommended by Task Forces that are comprised of acknowledged experts in the various structure determination methods. It is expected that these tools will improve both the efficiency and quality of annotation and help further promote data uniformity.

wwPDB members are: RCSB PDB (supported by NSF, NIGMS, DOE, NLM, NCI, NINDS and NIDDK), PDBe (Wellcome Trust, EU, BBSRC, NIH and EMBL), PDBj (BIRD-JST) and BMRB (NLM).
ChEBI, an Open-access Chemistry Resource for the Life Sciences: Facilities for On-line Submission and Curation

Marcus Ennis, Nico Adams, Adriano Dekker, Paula de Matos, Janna Hastings, Kenneth Haug, Duncan Hull, Zara Josephys, Gareth Owen, Steve Turner and Christoph Steinbeck

European Bioinformatics Institute, Wellcome Trust Genome Campus, Cambridge, UK

ChEBI (Chemical Entities of Biological Interest) is a database of ‘small’ molecular entities structured around a chemical ontology. It contains almost 600,000 entries, of which approximately 20,000 have been manually curated, as well as entries for groups (parts of molecular entities) and classes of entities. It provides a wide range of information such as chemical nomenclature, structures and related chemical values, and establishes interrelationships between entities in the ontology, in terms of both structure and role. ChEBI places a strong focus on quality, with exceptional efforts being applied to upholding IUPAC nomenclature recommendations and best IUPAC practices when drawing chemical structures.

To invite the community to participate more directly in the future growth and development of ChEBI, we have developed a web-based software utility to enable direct user submissions. Users are encouraged to carry out as much of their own manual curation as possible, e.g. by adding multiple synonyms and database cross-references, and by creating multiple relationships within the ontology. The submissions are automatically validated for uniqueness (both of name and chemical structure) and correctness (such as checking that no non-allowed cycles have inadvertently been created in the ontology graph structure, and that the ontology relationships which have been specified are allowed between entities of the relevant types). Once a submission has passed the required validations, it is submitted to the ChEBI database, at which time it receives its unique ChEBI identifier. It will then become visible to the public (as a preliminary entry) as part of the monthly ChEBI release. To date, ChEBI has received over 750 such external submissions.

Functional Annotation of Sugar Databases

Frederique Lisacek¹, Christine Hoogland¹, Elaine Mullen¹, Lynette Hirschman², Nicolle Packer³

¹Proteome Informatics Group, Swiss Institute of Bioinformatics, Geneva, Switzerland
²The MITRE Corporation, Mass., USA
³Biomolecular Frontiers Research Centre, Faculty of Science, Macquarie University, Sydney, Australia

Glyco-bioinformatics means different things to different people. It ranges, for example, from the mining of the genome for glycosylation pathway enzymes, through to analysis of microarray expression of glycosyltransferases, to glycoprotein identification data, to the development of tools that facilitate the interpretation of mass spectra and to the construction of structural glycan databases. Lastly, these data need to be mined so that the function of the glycan structures attached to the range of glycoconjugates in nature can be determined. The quality of the data in these databases needs to be maintained for this work to be meaningful.

For easier analysis, glycan structures are often characterised independent of the protein on which these sugars were originally attached. GlycosuiteDB (http://glycosuitedb.expasy.org/glycosuite/glycodb) is a freely accessible resource that has annotated this glycan structural data with protein and tissue information that sets the basis of functional interpretation. Information on cellular interactions associated with changes in glycosylation structures is essential for the understanding of the functional role of sugars and needs to be collected in order to exploit the knowledge of these posttranslational additions to proteins. One such database is SugarBind (http://sugarbind.expasy.org/sugarbind), which collates the known glycan binding sites of human pathogens. These structural glycan epitopes occur on both proteins and lipids and will add to the knowledgebase of glycosylation function and expand the structural data contained in GlycoSuiteDB. In this presentation, the knowledge obtained on the function of glycosylation by linking the information between two databases will be highlighted as an example of the value to be gained from manual curation of literature-extracted knowledge.
Glycome informatics has started to develop during this past decade, especially with the development of glycan structure databases by KEGG and others in the United States and Europe. However, with the development of various databases, many different formats representing glycan structures have emerged. Recently, the Consortium for Functional Glycomics (CFG) and EuroCarb have collaborated to develop a glycan structure standard for data exchange, based on the GLYDE-II xml representation for carbohydrates and glycoconjugates. In addition, web services have been developed in many glyco-databases to enable the exchange of such data. As a first attempt for glycan data standardization, an xml-based standard format was developed, along with several interfaces such that basic query searches can be implemented with a common standard interface. These interfaces include: retrieval of glycan information based on database ID, glycan substructure, mass and glycan composition. The retrieved data can also be used for operational queries such that more complex queries can be made. The implemented services in EuroCarb and CFG will in turn be registered in BioMoby such that workflows can be generated based on them. This will also potentially allow the incorporation of glyco-data with other related databases. Although this area of glyco-informatics is still in the early stages, many useful data resources and web services have been developed, and these will be presented.
Curation Of Protein Domain Models For The Conserved Domain Database (CDD)

Aron Marchler-Bauer, Stephen H. Bryant

National Center for Biotechnology Information, National Library of Medicine, 8600 Rockville Pike, Bethesda, MD 20894, USA

We present software applications, policies, and procedures that facilitate manual curation of protein domain models for the Conserved Domain Database (CDD). CDD is a collection of domain and full-length protein models assembled from a variety of sources. A subset of CDD has undergone thorough review and curation that focuses on (1) detecting and representing major events in a domain family’s evolutionary history, as these tend to coincide with divergence in function, (2) abstracting functional information and information on functional sites as available in the published literature, and (3) utilizing protein three-dimensional structure data to guide multiple sequence alignments via structure superposition and to annotate functional sites based on observed interactions with other molecules.

The current version of CDD contains 37407 alignment models, of which 5831 have been curated by NCBI. Together, they annotate about 74% of the protein sequences in the NCBI Entrez database.

InterPro Curation: Integrating Predictive Protein Signatures Into Biological Hierarchies

Sarah W Burge1, Louise Daugherty1, David Lonsdale1, Craig McAnulla1, Siew-Yit Yong1, Prudence Mutowo1, Amaia Sangrador1 and Sarah Hunter1.

1EMBL-EBI, The Wellcome Trust Genome Campus, Hinxton, CB10 1SD, UK

InterPro is an integrated database of predictive protein signatures used for the classification and automatic annotation of proteins and genomes. As InterPro curators, we are responsible for assimilating information from our member databases and communicating it to our end users in a way that adds value to each individual signature. We categorise signatures according to their type (for example, Family, Domain or Repeat) and annotate entries with links to other databases, abstracts and protein matches.

The InterPro database also identifies relationships between entries. For example, signatures at a general Family level are related to more specific subfamilies through a Parent/Child relationship. Families may also Contain individual Domains. In this manner, we aim to build up a hierarchy of InterPro entries that correctly represents relationships between biological families and domains. Users may then easily identify related proteins and signatures as the InterPro database attempts to map out biological hierarchies. Here we discuss InterPro relations, the criteria for their formation and how they may be useful to users. We will also discuss the challenges of representing biological hierarchies when automating relationship formation and the role manual curation plays in ensuring that we accurately represent biological networks.
Manual Biocuration, A UniProtKB/Swiss-Prot Perspective

Sylvain Poux1, Ioannis Xenarios1,2 and the UniProt consortium1,3,4

1Swiss Institute of Bioinformatics, Switzerland
2Vital-IT Group, Switzerland
3European Bioinformatics Institute (EBI), UK
4Protein Information Resource (PIR), USA

Manual curation of proteins is essential to provide high quality datasets that can then be programmatically parsed and allow the scientific community a rapid access to reliable information. Manually curated entries are stored in the UniProtKB/Swiss-Prot section of UniProtKB and constitute the highest priority of the UniProt consortium, with more than 60% of its staff being fully dedicated to this task.

High quality information is added by experienced biocurators, most of them having a strong background in research. Priority is given to the biocuration of proteins with impact on the largest number of users. Biocurators extract a maximum of pertinent information from the scientific literature. Information is reviewed, compiled, summarized and reported in the appropriate fields of a UniProtKB entry. Controlled vocabularies are used to structure the information of an entry and evidence tags are attached to each piece of information indicating its source. We also assign Gene Ontology (GO) terms to all UniProtKB entries during the biocuration process, by extracting information related to each of the GO ontologies with experimental evidence.

To avoid duplication of annotation efforts with other resources we establish collaborations with a number of databases (i.e. MODs) to ensure an accurate and mutually beneficial integration pipeline. Collaborations have started in fields such as proteomics and protein-protein interactions and future plans involve collaboration with the NCBI’s Genome project and the Reference Sequence (RefSeq).

This constant effort and the experience accumulated over the years has created a unique task force in the domain of manual biocuration.

ISA Software Suite: Supporting Standards-Compliant Experimental Annotation and Enabling Curation at the Community Level

Philippe Rocca-Serra1, Marco Brandizi2, Eamonn Maguire1, Nataliya Sklyar2, Chris Taylor2,3, Kimberly Begley1, Dawn Field1, Stephen Harris6, Winston Hide4, Oliver Hofmann4, Steffen Neumann6, Peter Sterk3,5, Weida Tong6 and Susanna-Assunta Sansone1

1Oxford e-Research Centre, University of Oxford, UK; 2The European Bioinformatics Institute, UK; 3Natural Environment Research Council, Environmental Bioinformatics Centre, UK; 4Harvard School of Public Health, USA; 5Genomic Standards Consortium, Wellcome Trust Sanger Institute, UK; 6US FDA, Center for Bioinformatics, National Center for Toxicological Research, USA; 7Leibniz Institute of Plant Biochemistry, Department of Stress- and Developmental Biology, Germany

Funders and journals require that researchers share their data, and encourage the enrichment and standardization of experimental metadata (1) (i.e., sample characteristics, technology/measurement types and instrument parameters, sample-to-data relationships) to make data sets comprehensible, reusable, underpinning future investigations. This situation necessitates better annotation at source (by data generators or community-based curation efforts), using software with automated content validation (2). Such software should support minimum information checklists and ontologies, such as those listed by the MIBBI Portal (3), and public ontology portals (i.e., http://www.ebi.ac.uk/ontology-lookup, http://bioportal.bioontology.org).

The Investigation/Study/Assay (ISA, 4) is the first general-purpose freely-available desktop software suite for curators and experimentalists designed to (i) regularize local management of experimental metadata, (ii) support community-defined ontology and reporting standards, and (iii) format studies for submission to public repositories.

Software, documentation, mailing lists, list of collaborators and case studies at: http://www.isa-tools.org

Ontology-based Tools to Enhance the Curation Workflow

Patricia L. Whetzel, Nigam H. Shah, Natasha F. Noy, Clement Jonquet, Cherie Youn, Paul Alexander, Michael Dorf, Mark A. Musen

Stanford University, USA

In order to effectively search, retrieve, and analyze data oftentimes it is curated and tagged with ontology terms. However, the amount of effort to curate the existing set of data resources is beyond the limits of purely manual curation. We present three ontology-based tools developed by the National Center for Biomedical Ontology to enhance the curation workflow: Ontology Widgets, Notes, and the Annotator. The Ontology Widgets provide a mechanism to use ontologies in Web-based forms without the need to locally parse and store the ontology. The widgets provide a variety of functionality including term autocompletion and ontology visualization. The Ontology Widgets are implemented for all BioPortal ontologies, including those from the OBO Foundry and Unified Medical Language System. The Notes feature of BioPortal allows structured term proposals to be submitted in order to request the addition or modification of a term in an ontology. The term proposals can be added directly via the BioPortal Web interface or programmatically via the Notes Web service. Notification of new Notes and replies are both RSS- and Email-enabled. Once the term curation process is complete, the OWL class or OBO stanza can be generated via the Notes Web service. Finally, the Annotator can be used to automatically process textual metadata to identify ontology terms found within the text. The Annotator can be accessed programmatically via the Annotator Web service and can be used with all BioPortal ontologies. In summary, the Ontology Widgets, Notes, and Annotator provide mechanisms to enhance curation by helping collect annotated data upon data submission, by facilitating ontology term curation, and by tagging unstructured textual data with ontology terms.

Gene Curation Software at the Rat Genome Database

Stan Laulederkind, Mary Shimoyama, Brad Taylor, Victoria Petri, Tim Lowry, Tom Hayman, Shur-Jen Wang, Jennifer Smith, Rajni Nigam, Jeff De Pons, Weisong Liu, George Kowalski, Melinda Dwinell, Diane Munzenmaier, Howard Jacob

Bioinformatics Program, Medical College of Wisconsin, 8701 Watertown Plank Rd, Milwaukee, WI, 53226, USA

At model organism databases data is curated for numerous biological categories using various ontologies and vocabularies. The Rat Genome Database (RGD) uses four different ontologies to standardize annotation information for genes and their associations with disease, phenotypes, and pathways. For manual gene curation this is all done in a single user interface of a web-based annotation tool developed at RGD. The same interface can be used for the curation of QTLs and strains. The development of the tool has been achieved through a collaboration of curators and software developers. Features have been tailored to the needs of the curators to allow optimum efficiency of the data entry portion of the curation process. Annotations using the Gene Ontology, Mammalian Phenotype Ontology, Pathway Ontology, and Disease Ontology can be done simultaneously in the same user interface. The search function of the tool has multiple options, giving the curators the opportunity to specify the search according to the immediate need. For ontology terms, gene, QTL, and strain objects the tool searches RGD data tables. For references the tool searches both internal and PubMed reference IDs. If an abstract reference is not already found at RGD, the tool automatically downloads the reference and assigns an RGD ID to it. Data objects, ontology terms, and references can be stored in multiples, providing various options for composing annotations. An editing step allows the composed annotation line to be altered before entering it in the database. The tool displays all current annotations for any gene that is selected. A second editing function is available by hyperlink for any pre-existing annotation displayed in the tool. The curation software has more than doubled literature curation production since replacing spreadsheet curation.
Data Curation Infrastructure In The iPlant Discovery Environment

Chris Jordan¹, Dan Stanzione¹, Doreen Ware²
¹The University of Texas at Austin, ²Cold Spring Harbor Laboratory

The iPlant Collaborative is a 5-year, National Science Foundation-funded effort to develop cyberinfrastructure to address a series of grand challenges in plant science. One of these grand challenges is the Genotype-to-Phenotype project, which seeks to provide tools, in the form of a web-based Discovery Environment, for understanding the developmental process from DNA to a full-grown plant. Addressing this challenge requires the integration of multiple data types that may be stored in multiple formats, with varying levels of standardization. Providing for reproducibility requires that detailed information be gathered documenting the experimental provenance of data, along with the computational transformations applied to data once it is brought into the iPlant environment. Handling the large quantities of data involved in high-throughput sequencing and other experimental sources of bioinformatics data requires a robust infrastructure for storing and reusing large data objects. We describe the currently planned workflows to be developed for the Genotype-to-Phenotype discovery environment, the data types and formats that must be imported and manipulated within the environment, and the data model developed to address this complexity of data types and formats. We then describe in detail the new features being added to the Discovery Environment to enable and encourage collective curation of bioinformatics datasets, in particular the model for capturing and enhancing provenance metadata as new data elements are brought into the iPlant discovery environment, and the capabilities for interaction with reference database, focusing not just on the ability to retrieve data from such data sources, but on the ability to use the iPlant Discovery Environment to further populate these important resources.

Application of Next-Generation Sequence Data For Rice Genome Analyses

Yoshihiro Kawahara
National Institute of Agrobiological Sciences, Japan

In recent years, emerging techniques of next-generation sequencing (NGS) have been used for various purposes: genome re-sequencing, detection of SNPs and structural variations, transcriptome analysis by RNA-seq, etc. While the cost and time required for obtaining genome and transcriptome sequences have drastically reduced, computational processes of large-scale data became time-consuming, so that elaborated bioinformatics methods are needed. Here we present our activities about genome re-sequencing of japonica rice, and transcriptome analyses to find agronomically useful genes. For biodiversity studies of cultivars whose nucleotide differences are quite small, an extremely high-quality reference genome is needed. Therefore, we attempted to improve the genome sequence of a japonica cultivar Nipponbare using Illumina and 454 data. Mapping of more than 100 million reads of Illumina and 454 detected ~1-10 X 10^{-5} errors per site. In addition, approximately 1,700 SNPs were found between two individuals. Furthermore, to add to knowledge about functions of the rice genes, we conducted mRNA-seq analyses under several conditions using Illumina data. In an analysis of salinity stress, we compared expression levels of each transcript on the RPKM (reads per kilobase of exon per million mapped sequence reads) basis before and after the stress. As a result, we found that a significant number of genes are induced by salinity stress, although many of their biological functions are currently unknown. While DNA microarray studies have a relatively long history, an advantage of mRNA-seq analyses is that it can predict novel exon-intron structures and can estimate expression levels of such unannotated transcripts without any prior knowledge. So far, we could predict ~1,000 unannotated transcript structures by the TopHat and Cufflinks programs, and found that some of them are stress-inducible. We are currently planning to extend this research to some other conditions and cultivars. Finally, we present our newly developed viewer for NGS data, which will be a part of the Rice Annotation Project Database.
MAIZEGDB.ORG, the Maize Genetics Cooperation and the 2500 MB B73 Genome-Generated Tsunami

Mary Schaeffer*1,2, Jack Gardiner*1,5, Lisa Harper*4, Carson Andorf5, Darwin Campbell1, Ethy Cannon5, Taner Sen1,5, Carolyn Lawrence §1,5

1USDA-ARS USA, 2University of Missouri USA, 3University of Arizona USA, 4UC Berkeley USA, 5Iowa State University USA, *Curator, §Director

Advances in sequencing technology have made it possible to sequence the 2500 MB B73 maize genome, both cheaply and in a relatively short time. Nearly simultaneously, other sequencing-based data are on the leading edge of a data tsunami: sequenced differences (currently >300,000 SNP for >1000 inbred lines and related species) and gene expression data. The MaizeGDB team integrates access to these datasets after considerable cooperation among both data users and suppliers, and typically while the data are being generated. In this way, data suppliers have resources in place to format data ready for integration with MaizeGDB. And, we learn what details, including links to other resources, are needed to support our user community and what interface development is required. An interesting result is that we inadvertently serve as outreach for the data suppliers, explaining the caveats and expected availabilities of data not yet at MaizeGDB. Contact with users is at conferences, by online tutorials, personal visits, or by surveys from the Maize Executive Genetics Committee addressed to all persons listed in MaizeGDB as a Cooperator.

Arabidopsis thaliana: further Exploiting This Plant Reference Genome

Mara Sangiovanni1, Alessandra Vigilante1 and Maria Luisa Chiusano1

1Department of Soil, Plant, Environmental and Animal Production Sciences, University of Naples Federico II, Via Università 100, 80055 Portici, Italy

Arabidopsis thaliana is the reference in plant genomics since its complete genome sequence was the first one to be made available in 2000. As a reference model the Arabidopsis genome should be fully reliable and safely annotated; moreover its organization should be well understood in terms of evolutionary mechanisms that gave rise to the actual genome structure. However the presence of widespread intragenome duplications, together with the loss of many gene copies, associated to possible ancient and recent polyploidization events, really complicates the interpretation of the factors contributing to the genome shaping, thus limiting also the role of this genome as a reference in plant comparative genomics.

To further exploit the information available, and with the aim of supporting the genome annotation of other plant and crop species, we investigated the organization of the Arabidopsis genome in terms of paralogous genes. We identified all the possible pair-wise similarities between genes classifying structurally related ones into networks, with each gene belonging to only one network given the presence of one or more paralogy relationships. The organization of the genome into networks of duplicated genes therefore provides a novel view for intra-genome and inter-genome comparative analyses, also permitting an appropriate investigation of gene families evolution. We also focused on the identification of single copy genes (singletons), because their presence in a highly duplicated genome is still an intriguing evolutionary issue. Furthermore the analysis of singleton genes revealed some flaws into the annotation data of Arabidopsis genome further indicating the need of a more accurate annotation process.

This work is supported by the Agronanotech Project (Ministry of Agriculture, Italy)
Quantitative trait loci (QTL) analysis is an important approach in elucidating the genetic basis underlying complex traits. However, as entire genomes are being sequenced and an increasing amount of genetic and expression data is being generated, a challenge remains: linking phenotypic variation to the underlying genomic variation. Common practice in major biological databases is to curate QTL coordinates from published literature. At the Sol Genomics Network (SGN) (http://solgenomics.net), we have developed a QTL module (http://solgenomics.net/qtl/) which allows QTL researchers to upload their raw genotype and phenotype data, set statistical parameters and perform on-the-fly QTL analysis using user-friendly web interfaces. The tool implements R/QTL (http://www.rqtl.org), an add-on package for the R statistical software. In-house and open source Perl algorithms are used for on the web visualization of the QTL mapping output and cross-linking it to relevant genetic, expression, and genomic datasets at SGN and external databases, using SGN’s Comparative Map Viewer and GBrowse. This data integration and synthesis can facilitate identification of candidate genes, elucidation of the underlying variation at the molecular level and application of marker assisted breeding for a trait of interest. Currently, the QTL mapping is implemented for backcross and F2 diploid QTL populations and for traits with continuous variation. It allows researchers to explore and analyze genetic, expression and genomic data relevant to pre-publication QTLs. The QTL tool is freely available on SGN.
Hyperlink Management System for Creating Maintenance-Free Hyperlinks among Major Biological Databases

Tadashi Imanishi

BIRC, AIST, Japan

Hyperlink Management System (HMS) is a system for automatically updating and maintaining hyperlinks among major public databases in the field of life science. We daily create corresponding tables of data IDs of major biological databases for human and mouse molecules, and provide a CGI-program that returns correct and up-to-date URLs for showing data of various databases that correspond to user-specified IDs. The HMS can deal with various IDs for human: accession numbers of International Nucleotide Sequence Databases (INSD), HUGO Gene Symbols, Gene IDs of NCBI, UniProt accession numbers, PDB IDs, HIT and HIX of H-InvDB, and many others, and it can return URLs of various databases: H-InvDB, HUGO Gene Nomenclature Committee Database, NCBI Entrez Gene, UniProt, PDB, OMIM, PubMed, and others. For example, more than 23,000 pages of Locus view of H-InvDB are reachable by using HUGO Gene Symbols through the HMS. For mouse molecules, we collect IDs of INSD, Ensembl, UniProt, KEGG, MGI, IKMC, and FANTOM clones. Furthermore, we recently started dealing with chemical compounds and drugs, collecting IDs of DrugBank, KEGG Drug, PubChem, ChEBI, PDBChem, KEGG Compound, and CAS. Not only the CGI-program, the HMS provides a Web page for finding and opening URLs of these databases. Although hyperlinking is an effective way of relating biological data among different databases, updating hyperlinks has been a laborious work. The HMS fully automates the job, enabling maintenance-free hyperlinks. We also developed the ID Converter System (ICS) for simply converting data IDs by using corresponding tables in the HMS. The HMS and IDS are freely available at http://biodb.jp/.

Semantic Encoding of Complex Information

Steven Vercruysse, Martin Kuiper

Systems Biology group, NTNU (Norwegian University of Science and Technology), Norway

The biosciences are a domain where pieces of information are often highly complex and composite in nature, and where the specific context is often crucial for a correct understanding of a biological system. We took it as a challenge to develop a flexible curation approach that allows a full semantic description of any type of information. Current approaches mainly focus on a table or spreadsheet-based formatting of standardized facts. However, in practice curators often encounter extra relevant context that can not be captured in these predesigned templates. This becomes especially apparent when curating a variety of detailed facts from diverse publications. Also controlled languages only allow partial capture of this variety of facts and context. Therefore we designed a new, general method to capture composite information into a flexible, yet semantically clear format.

We present a flexible method to graphically indicate the structure in any linear series of controlled terms (linked to semantic identifiers). For this, we go beyond representing information with only triples (like RDF); we describe a method to generate structured tuples of arbitrary complexity and depth. We illustrate how most linguistic structures can be built from only three simple connector types (with one being the most prevalent), and how curators can use these 'Visual Syntax Markup' (VSM) connectors to visually indicate the implied structure of any curated sentence. Our approach makes it possible to encode information that is as complex as what natural language presents. We describe the VSM method, its semantics, and its possible application in a user interface. We furthermore show how it may serve as a basis that unites a diverse set of curation initiatives and input forms.
NamesforLife Semantic Resolution Services for the Life Sciences: Moving Towards an Extensible and Interoperable System for Naming

George M Garrity1,2, Charles Parker1, Dorothea Taylor1, Kara Mannor1, Catherine Lyons1

1NamesforLife, LLC, USA and UK
2Michigan State University, USA

A major challenge in bioinformatics, life sciences, and medicine is using correct and informative names. While this sounds simple enough, many different naming conventions exist in the life sciences and medicine that may be either complementary or competitive with other naming conventions. For a variety of reasons, proper names are not always used, leading to an accumulated semantic ambiguity that readers of the literature and end users of databases are left to resolve on their own. This ambiguity is a growing problem and the biocuration community is aware of its consequences.

To assist those confronted with ambiguous names (which not only includes researchers but clinicians, manufacturers, patent attorneys, and others who use biological data in their routine work), we developed a generalizable semantic model that represents names, concepts, and exemplars (representations of biological entities) as distinct objects. By identifying each object with a Digital Object Identifier (DOI), it becomes possible to place forward-pointing links in the published literature, in databases, and vector graphics that can be used as part of a mechanism for resolving ambiguities, thereby “future proofing” a nomenclature or terminology. A full implementation of the N4L model for the Bacteria and Archaea was released in April and encompasses 14264 names, 13831 taxonomic concepts and 13892 exemplars. It is backed by 11456 distinct references. The system is professionally curated and represents a Tier III resource in Parkhill’s view of informatic services. A variety of tools and web services have been developed for readers, publishers, and others and we are incorporating other taxonomies into the N4L data model, as well as adding additional phenotypic, genotypic, and genomic information to the existing exemplars to add greater value to end users.

Bio2RDF: Convert, Provide And Reuse.

Marc-Alexandre Nolin1, Jacques Corbeil1, Luc Lamontagne1, Michel Dumontier1,2

1Laval University, Canada
2Carleton University, Canada

The Bio2RDF project uses open-source Semantic Web technologies to provide interlinked life science data in order to maximize productivity and facilitate biological knowledge discovery. Using both syntactic and semantic data integration techniques, Bio2RDF puts into practice a simple methodology to generate and seamlessly integrate machine-interpretable data that can be powerfully interrogated with SPARQL-based queries to answer sophisticated questions.

At its core, database records are converted into a set of statements or so-called triples that are captured together as a named graph annotated with provenance. The records and the entities they are about are provided with a Uniform Resource Identifier (URI) of the form http://bio2rdf.org/prefix:identifier, where the prefix indicates a reserved name for the dataset, record or terminological resource. The application of this simple method allows resources from over 40 datasets to integrate seamlessly at the syntactic level irrespective of whether the original data contains non-Bio2RDF URIs.

However, when original data providers such as Uniprot provide their own RDF they will rightfully use URIs that resolve to their servers, but what should they do for externally defined entities? If they follow in Bio2RDF’s footsteps then every data provider will use a different URI. However, should original data providers present and implement a URI scheme, then it becomes possible for others to establish stable links to their resources. As such, we will witness the birth of a more stable linked data network, ensuring that data providers need not provide third party data in a redundant manner.

Tetsuro Toyoda1.

1 Bioinformatics And Systems Engineering (BASE) division, RIKEN, Japan

Recent life science research activities are characterised by the conduction of research through large-scale databases in global communications with researchers at remote locations. In order to realize virtual laboratories in which researchers exchange large scale data and opinions to conduct research together, we have developed RIKEN SciNetS (Scientists' Networking System) as a Cloud information infrastructure. The system is designed to construct myriad highly-secured virtual laboratories with essential functions for international collaborative research such as a powerful database construction function technologies, an electric lab book function which enables control of confidential information, and a message function, and those functions have been developed on the bases of the Semantic Web technologies.

The life sciences databases published by RIKEN and others of various species including mammalians and plants are constructed in SciNetS as virtual laboratories and related data records are connected by semantic links over the laboratories; therefore, an integrated database over various omics data records is realized. The system allows users to browse such integrated data in HTML table format, search data by specifying keywords using semantic inference search engine called GRASE, and download a data set in various standardized formats.

Virtual laboratories published on RIKEN SciNetS are accessible by common web browser at http://database.riken.jp.


TogoDB + TogoWS: a data integration platform for the Semantic Web

Toshiaki Katayama1, Mitsuteru Nakao2

1 Human Genome Center, Institute of Medical Science, University of Tokyo, Japan
2 DataBase Center for Life Science, Research Organization of Information and Systems, Japan

The Database Center for Life Science (DBCLS) is established to provide integrated biomedical resources. Our group has focused on improving the interoperability of bioinformatics web services and developing technologies for integration of heterogeneous life science data. Web services have become widely used in bioinformatics analysis, but there exist incompatibilities in interfaces and data types, which prevent users from fully utilizing a combination of these services. In this situation, we developed the TogoWS system, which enables users to access public web services including DDBJ, PDBj, KEGG, EBI and NCBI through a unified REST API. The API provides intuitive URIs to search, retrieve, parse and convert the database entries and which eliminates incompatibilities in those services. Each database entry can be specified by a unique URI, but more importantly, even a part of the entry (in a designated data format, such as XML, JSON and RDF) has a specific URI. This concept fits very well with the Semantic Web because every object in RDF data needs to be represented as a unique and persistent URI. Nowadays, rapid accumulation of data made it hard to obtain all related information stored in the distributed databases without having semantic metadata. Therefore, the Semantic Web is being considered as a promising technology to integrate a huge amount of heterogeneous data in life science. To explore the feasibility of this technology, we organized the DBCLS BioHackathon, an international workshop, gathering database providers and service developers around the world. As a result, we released the Tokyo manifesto which encourages database providers to expose their data in RDF format with officially authorized URIs, and service providers to respect those primary URIs in their applications. We are then developing our in-house database system, TogoDB, to export its contents with semantic annotations through the TogoWS service as a part of the Linked Data.
Poster Abstracts

Presenting authors are requested to stay at their posters during the following hours.

**Odd numbers:** October 12(Tue), 17:30-18:30; October 13(Wed), 18:30-19:30

**Even numbers:** October 12(Tue), 18:30-19:30; October 13(Wed), 17:30-18:30

### P1 Text Mining

**P1-1** Apply of Textmining Method to Study the Roles in Improving the Health by Lactoferrin, a Multi-Functional Milk Protein  
Kei-ichi Shimazaki

**P1-2** Combining Support Vector Machines (SVMs) with Textpresso Category Searches for Improved Data Type Curation Efficiency at WormBase  
Kimberly Van Auken

**P1-3** An Open Resource for Collecting and Distributing Textual Evidences for Protein Annotation  
Xinghua Lu

**P1-4** ODIN: An Advanced Interface for the Curation of Biomedical Literature  
Fabio Rinaldi

### P2 Literature Curation

**P2-1** Annotating Whole-genome Sequencing in the Catalogue of Somatic Mutations in Cancer  
Chai Yin Kok

**P2-2** Linking Phenotype Data to the Genome: RGD a Unique Resource for Rat Strains and QTLs  
Rajni Nigam

**P2-3** Logical Operation Based Literature Association with Genes and its application, PosMed  
Yuko Makita

**P2-4** Epigenetics Knowledgebase-Role of Epigenetic alterations in Cancer  
Priti Y. Ramamohan

**P2-5** Clipro-A Comprehensive Knowledgebase of Clinically Significant Proteins in Biofluids  
Jaya Iyer

**P2-6** Challenges to Extract Kinetic Data from Literature  
Ulrike Wittig

**P2-7** Manual Curation of Vertebrate Proteins in UniProtKB  
Michael J. Gardner

**P2-8** Manual Biocuration, A UniProtKB/Swiss-Prot Perspective  
Sylvain Poux

**P2-9** Gene Ontology Annotation of Hearing  
Guvanch Ovezmyradov

**P2-10** caNanoLab - A Tool To Benefit Biomedical Nanomaterials Research  
Michal Lijowski
<table>
<thead>
<tr>
<th>Page</th>
<th>Title</th>
<th>Author</th>
</tr>
</thead>
<tbody>
<tr>
<td>P2-11</td>
<td>ACBD: Database for Ascidian Chemical Genomics</td>
<td>Yuichiro Hira</td>
</tr>
<tr>
<td>P2-12</td>
<td>Towards Automated Annotation of Ortholog Clusters Using Natural Language Processing Technologies</td>
<td>Yasunori Yamamoto</td>
</tr>
<tr>
<td>P2-13</td>
<td>Information on Transcriptional Regulation and Signal Transduction of Escherichia coli K-12 Integrated in the Database RegulonDB</td>
<td>Socorro Gama-Castro</td>
</tr>
<tr>
<td>P2-14</td>
<td>The BioGRID Interaction Database</td>
<td>Michael S. Livstone</td>
</tr>
</tbody>
</table>

**P3 Ontologies**

<table>
<thead>
<tr>
<th>Page</th>
<th>Title</th>
<th>Author</th>
</tr>
</thead>
<tbody>
<tr>
<td>P3-1</td>
<td>Ontology on Cell Lines</td>
<td>Usha Mahadevan</td>
</tr>
<tr>
<td>P3-2</td>
<td>eagle-i: An Ontology-Driven Framework for Biomedical Resource Curation and Discovery</td>
<td>Erik Segerdell</td>
</tr>
<tr>
<td>P3-3</td>
<td>Using Multiple Ontologies to Curate Phenotype Data</td>
<td>Mary Shimoyama</td>
</tr>
<tr>
<td>P3-4</td>
<td>Disease Ontology: Logical Definitions and Relations</td>
<td>LaRon Hughes</td>
</tr>
</tbody>
</table>

**P4 Genome Sequence Annotation**

<table>
<thead>
<tr>
<th>Page</th>
<th>Title</th>
<th>Author</th>
</tr>
</thead>
<tbody>
<tr>
<td>P4-1</td>
<td>Microbial Genome Annotation at NITE and Annotator-Oriented Development of Annotation Workbench</td>
<td>Natsuko Ichikawa</td>
</tr>
<tr>
<td>P4-2</td>
<td>ENA as an Information Hub</td>
<td>Petra ten Hoopen</td>
</tr>
<tr>
<td>P4-3</td>
<td>Gene Trek in Prokaryote Space (GTPS) Reduces GIGO in the Annotation of Microbial Genome Sequences</td>
<td>Takehide Kosuge</td>
</tr>
<tr>
<td>P4-4</td>
<td>Automatic Protein Clustering as a Basis of Automatic Annotation</td>
<td>Naoki Sato</td>
</tr>
<tr>
<td>P4-5</td>
<td>The Challenge of Eukaryote Genome Annotation at Genoscope</td>
<td>Betina M. Porcel</td>
</tr>
<tr>
<td>P4-6</td>
<td>Annotation Strategies for High-Coverage, Low-Coverage and Next-Generation Sequencing Assemblies in Ensembl</td>
<td>Jan Vogel</td>
</tr>
<tr>
<td>P4-7</td>
<td>Extension of Database of Transcription Start Sites (DBTSS) with a Large Number of Sequences and Analysis of Bi-directional Promoters</td>
<td>Riu Yamashita</td>
</tr>
<tr>
<td>P4-8</td>
<td>The Hymenoptera Genome Database</td>
<td>Monica Cecilia Munoz-Torres</td>
</tr>
</tbody>
</table>
### P4-9  A Trial for Genome Re-annotation Using the Literature and Annotation Reference Information
Takatomo Fujisawa

### P5 Comparative Genomics

**P5-1**  
*Analyzing Inverse Symmetry with Original and Terminal Sites of Prokaryotic Genomes*  
Tzu-Ting Hsu

**P5-2**  
*An Italian SOLanaceae Integrated Platform (ISOL@): the Kick Start to Exploit Solanaceae Comparative Genomics in Solanaceae*  
Alessandra Traini

**P5-3**  
*H-DBAS: Human-transcriptome Database for Alternative Splicing, Version 5.0*  
Jun-ichi Takeda

**P5-4**  
Cancelled

**P5-5**  
*Evolutionary Analysis of Nucleotide Sequences Using G-compass, a Web-based Comparative Genome Browser between Human and Vertebrates*  
Yuichiro Hara

**P5-6**  
*Evola: A Database of Orthologous Genes and Gene Families between Human and Other Vertebrates Detected by Whole Transcriptome Analyses*  
Ryuichi Sakate

### P6 Functional Genomics

**P6-1**  
*Development of 12 Genic SSR Markers for a Biofuel Grass, Miscanthus sinensis (Poaceae)*  
Chi-Chun Huang

**P6-2**  
*Profile on Human Druggable Space Revealed by Integrated Analyses of Bioinformatic and Chemoinformatic Resources*  
Yasuyuki Fujii

### P7 Crop/Plant Genome Curation

**P7-1**  
*Arabidopsis thaliana: Further Exploiting This Plant Reference Genome*  
Mara Sangiovanni

**P7-2**  
*MAIZEGDB.ORG, the Maize Genetics Cooperation and the 2500 MB B73 Genome-Generated Tsunami*  
Mary L. Schaeffer

**P7-3**  
*Rice Genes in Oryzabase*  
Yukiko Yamazaki

### P8 Community Curation

**P8-1**  
*ChEBI, an Open-access Chemistry Resource for the Life Sciences: Facilities for On-line Submission and Curation*  
Marcus Ennis
<table>
<thead>
<tr>
<th>Session</th>
<th>Title</th>
<th>Author(s)</th>
</tr>
</thead>
<tbody>
<tr>
<td>P8-2</td>
<td>Knowledge Discovery and Database Construction for Genes Contributing to &quot;Sustainable World&quot; and &quot;Human Health&quot; &amp; tRNADB-CE under Collaboration with Undergraduate Students and Senior Scientists</td>
<td>Hiroshi Uehara</td>
</tr>
<tr>
<td>P8-3</td>
<td>Incorporating Community Annotation Interfaces into the CIPRO2.5 Database with Comprehensible Sketches to Support Quick Annotations of Proteome Data</td>
<td>Keisuke Ueno</td>
</tr>
<tr>
<td>P8-4</td>
<td>CIPRO 2.5: Ciona intestinalis Protein Database, a Graphically Integrated Proteome Database with Facility for Biocuration</td>
<td>Toshinori Endo</td>
</tr>
<tr>
<td>P9-1</td>
<td>H-InvDB: A Comprehensive Annotation Resource for Human Transcriptome</td>
<td>Chisato Yamasaki</td>
</tr>
<tr>
<td>P9-2</td>
<td>Data Integration Model and GUIs Used in Human Genome Network Platform</td>
<td>Kazuho Ikeo</td>
</tr>
<tr>
<td>P9-3</td>
<td>POPcorn: A PrOject Portal for corn</td>
<td>Taner Z. Sen</td>
</tr>
<tr>
<td>P9-4</td>
<td>Towards Developing Common Standards for Genome Sequence and Annotation</td>
<td>Tatiana Tatusova</td>
</tr>
<tr>
<td>P9-5</td>
<td>Literature Curation of Protein Interactions: Measuring Agreement across Major Public Databases</td>
<td>Andrei L. Turinsky</td>
</tr>
<tr>
<td>P9-6</td>
<td>Neuroscience Resources and the Neuroscience Information Framework</td>
<td>Maryann Martone</td>
</tr>
<tr>
<td>P9-7</td>
<td>Data Standardization for Glyco-informatics</td>
<td>Kiyoko F. Aoki-Kinoshita</td>
</tr>
<tr>
<td>P9-8</td>
<td>SciNeS Search: Inference Search over an Integrated Life-sciences Database Based on the Semantic Web</td>
<td>Norio Kobayashi</td>
</tr>
<tr>
<td>P9-9</td>
<td>Metabolic Pathway Integration of Vibrio vulnificus CMCP6</td>
<td>Kwangjoong Jeong</td>
</tr>
<tr>
<td>P9-10</td>
<td>Challenges and Solutions to Rice Biological Data Curation at IRRI</td>
<td>Richard Bruskiewich</td>
</tr>
<tr>
<td>P9-11</td>
<td>Integral Presentation of Experimental Results and Bioinformatic Analyses on Ciona intestinalis Proteins</td>
<td>Kouki Yonezawa</td>
</tr>
<tr>
<td>P9-12</td>
<td>The Neuroscience Information Framework: A Unified Semantic Framework for Discovery and Integration of Biomedical Data and Resources on the Web</td>
<td>Jeffrey S. Grethe</td>
</tr>
</tbody>
</table>
### P10 Curation Tools/Workflows

<table>
<thead>
<tr>
<th>P10-1</th>
<th>WormBase Curation Interfaces and Tools</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Xiaodong Wang</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>P10-2</th>
<th>RDF Curator: A Novel Workflow that Generates Semantic Graph from Literature for Curation Using Text Mining</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Yusuke Komiyama</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>P10-3</th>
<th>InterPro Curation: Integrating Predictive Protein Signatures into Biological Hierarchies</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Sarah Burge</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>P10-4</th>
<th>Development of Cooperative Work Tools and Workflow for Multimodal Genome Annotation</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Shinobu Okamoto</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>P10-5</th>
<th>Ontology-based Tools to Enhance the Curation Workflow</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Patricia L. Whetzel</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>P10-6</th>
<th>The Annotation of Olfactory Receptor Gene Family in H-InvDB</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Akiko Ogura-Noda</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>P10-7</th>
<th>DICHOT - An Automatic Annotation System for Folding and Unfolding Regions of Proteins</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Satoshi Fukuchi</td>
</tr>
</tbody>
</table>

### P11 Data Visualization

<table>
<thead>
<tr>
<th>P11-1</th>
<th>EMAGE: A Database of Spatially Integrated Gene Expression Data</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Shanmugasundaram Venkataraman</td>
</tr>
</tbody>
</table>

### P12 Journal/Database Interactions and Collaborations

<table>
<thead>
<tr>
<th>P12-1</th>
<th>Allergens in ChEBI - Collaboration with the IEDB</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Zara Josephs</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>P12-2</th>
<th>The International Molecular Exchange Consortium</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Sandra Orchard</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>P12-3</th>
<th>MEDALS: METI Database Portal for Life Science and Renewal-Checker to Keep up to Date Information</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Akihiro Matsuya</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>P12-4</th>
<th>Mapping Bio-databases and Software by Text-mining</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Kentarou Mamiya</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>P12-5</th>
<th>NamesforLife Semantic Resolution Services for the Life Sciences: Moving Towards an Extensible and Interoperable System for Naming</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>George M. Garrity</td>
</tr>
</tbody>
</table>

### P13 Next Generation Sequence Annotation

<table>
<thead>
<tr>
<th>P13-1</th>
<th>454 RNA-Seq Data Analysis: What is New and What is Unchanged?</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Nunzio D'Agostino</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>P13-2</th>
<th>Analysis of Oryza nivara Genome via Illumina Sequence Platform</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Cheng-Chieh Wu</td>
</tr>
</tbody>
</table>
### P13-3  
*Towards the Draft Genome Sequences of Wild Rice Species: the Landmarks for Future Oryza Genomics*  
Hajime Ohyanagi

### P14 Next Generation Curation

**P14-1**  *RefEx: Reference Expression Dataset for Functional Curation of Transcriptomes*  
Hiromasa Ono

**P14-2**  *Semantic Encoding of Complex Information*  
Steven Vercruysse

### P15 Others

**P15-1**  *DDBJ Read Archive / DDBJ Omics Archive*  
Yuichi Kodama

**P15-2**  *PubMedScan: A Keyword-free Recommender of PubMed Articles*  
Katsuhiko Murakami

**P15-3**  *UniRule - Automatic Annotation in UniProtKB*  
Alan James Bridge

**P15-4**  *Rebuilding the Card Catalog - Resolving the Problem of Orphan Enzyme Activities*  
Alexander Glennon Shearer

**P15-5**  *An Automatic Method for Large-scale Functional Annotations Based on a Tree Alignment Algorithm*  
Masumi Itoh

**P15-6**  *HEAT: A New Tool for Gene Set Enrichment Analysis Using Comprehensive Annotation of Human Genes in H-InvDB*  
Miho Sera

**P15-7**  *Human Transcript Database Search Showed Existence of Extremely Short Introns*  
Makoto K. Shimada

**P15-8**  *The Latest Information and Perspective of Human Gene and Protein Database (HGPD) as Human Proteome Study*  
Yukio Maruyama

**P15-9**  *The Bioinformatics Links Directory: A Community Curated Collection of Bioinformatics Links, Tools and Databases*  
B. F. Francis Ouellette
Apply of Textmining Method to Study the Roles in Improving the Health by Lactoferrin, a Multi-Functional Milk Protein

Kei-ichi SHIMAZAKI1, Tatsuya KUSHIDA2, Koji YAMAUCHI3 and Toshihisa TAKAGI4

1Hokkaido University, Sapporo, 2Nalapro Technologies, Inc., Tokyo 3Food Science & Technology Institute, Morinage Milk Co., Ltd., Zama, 4Database Center for Life Science, ROIS, Tokyo

Lactoferrin is a metal-binding glycoprotein found in milk, blood and other exocrine secretions. This is a multi-functional protein that exhibits many activities such as: anti-microbial, anti-viral, immunomodulatory, anti-inflammatory, anti-tumor, anti-metastatic, cell growth-promoting, and anti-oxidant activities, as well as regulation of granulopoiesis and iron absorption, etc. To date, a number of academic reports concerning the biological activities of lactoferrin have been published and are easily accessible through public databases. In order to overcome the information overload associated with lactoferrin information, we have applied the text mining method to the accumulated lactoferrin literature. To this end, we used the information extraction system GENPAC (provided by Nalapro Technologies Inc., Tokyo), which uses natural language processing and text mining technology. Using GENPAC, text extraction was carried out on literature containing the term “lactoferrin” and any of keywords concerning health conditions or diseases from PubMed. Subsequently, network mappings of the information obtained were produced using Cytoscape. We will exhibit that such textmining method and information visualization analysis is useful in studying novel relationships among a multitude of lactoferrin functions and mechanisms to improve our health.

Combining Support Vector Machines (SVMs) with Textpresso Category Searches for Improved Data Type Curation Efficiency at WormBase

Kimberly Van Auken1, Ruihua Fang1, Juancarlos Chan1, Hans-Michael Müller2, and Paul Sternberg1,2

1WormBase, 2Textpresso, Division of Biology, California Institute of Technology, Pasadena, CA USA

Identifying relevant documents for curation and extracting key facts from the literature are the two main tasks of biological database curators. For many databases, these two tasks are still performed manually and thus, can be inefficient and insufficient for coping with large corpora. The Textpresso (http://textpresso.org) text mining and information extraction system employs keyword and category searches to identify biological facts from the full text of publications. Previously, we described Textpresso category searches that aid in the curation of Gene Ontology Cellular Component annotations1. Although Textpresso searches identify potentially curatable papers with only 62% precision, the precision in actually extracting the curatable information is very high (97%). To improve the precision of this first step in the curation pipeline, we tested the effectiveness of including a filtering step in which the results of an SVM model that predicts which papers are most likely to include expression data are combined with Textpresso category searches to potentially produce a smaller set of papers from which annotations could be made. Analysis of 50 new research articles acquired by WormBase (http://www.wormbase.org) indicates that the SVM model is sufficient for filtering 75% of the false positive documents returned by Textpresso searches alone. We are extending this combinatorial curation pipeline to additional data types, namely Gene Ontology Molecular Function curation. Our initial results indicate that a combined SVM and Textpresso category search can filter up to 76% of false positive papers for the subset of Molecular Functions that includes protein and nucleic acid interactions. Combining SVMs with Textpresso searches thus improves curation efficiency by reducing the number of papers curators must examine to make high quality, experimentally based annotations.

In the biomedical text-mining domain, one of the most challenging aspects of literature-based protein annotation is to extract from biomedical texts the salient textual regions describing biological concepts, which can be further utilized, instead of the whole text, as textual evidences to support gene and protein annotation in both manual and automatic ways. Although the biomedical articles, such as MEDLINE abstracts from PubMed, associated with multiple Gene Ontology annotations are readily available, there are few if any available documents in which text regions are segmented and labeled with specific annotations. The lack of such training corpus causes the research in this area (i.e. automatically extracting information from biomedical literature for protein annotation) remains sparse. Therefore, there are needs to develop a resource for collecting and distributing textual evidence for protein annotation. We collected a corpus of MEDLINE article abstracts with specified requirements and introduced a resource to facilitate domain experts manually extracting regions, e.g. sentences, from the corpus as textual evidences for protein GO annotation. The region-tagged corpus could be used as a training set for automatic GO annotation.

ODIN: An Advanced Interface for the Curation of Biomedical Literature

Fabio Rinaldi¹, Simon Clematide¹, Gerold Schneider¹, Martin Romacker², Thérèse Vachon²

¹University of Zurich, Switzerland
²NITAS/TMS, Text Mining Services, Novartis AG, Switzerland

We present ODIN (Ontogene Document INspector): a system for interactive curation of biomedical literature, developed within the scope of the SASEBio project, as a collaboration between the OntoGene group at the University of Zurich and the NITAS/TMS group of Novartis Pharma AG. The purpose of the system is to allow a human annotator/curator to leverage upon the results of an advanced text mining system in order to enhance the speed and effectiveness of the annotation process. The OntoGene system takes as input a document (e.g. a full paper from PubMed Central) and processes it with a custom NLP pipeline, which includes Named Entity recognition and relation extraction. Entities which are currently supported include proteins, genes, experimental methods, cell lines, species. Entities detected in the input document are disambiguated with respect to a reference database (UniProt, EntrezGene, NCBI taxonomy, PSI-MI ontology). The annotated documents are handed back to the ODIN interface, which allows multiple display modalities. The curator/annotator can view the whole document with in-line annotations highlighted, or can browse the extracted entities and be pointed back to the mentions of the entities within the original document. All entity mentions are entirely editable: the curator can easily add or delete any of them, and also change their extent (i.e. add/remove words to its right or left) with a simple click of the mouse. Different entity views are supported, with sorting capabilities according to different criteria (entity type, entity mention, confidence score, etc.). Selective highlighting of text units (e.g. sentences containing desired entities) is supported. Additionally, extensive logging functionalities are provided. All documents and entities are fully interlinked to reference databases, for the purpose of simplified inspection. Entities can be grouped in classes (e.g. by species) and actions can be applied to whole classes, for selective editing or removal.
Annotating Whole-genome Sequencing in the Catalogue Of Somatic Mutations In Cancer

C Y Kok1, S A Forbes1, N Bindal1, S Bamford1, C G Cole1, M Jia1, D Beare1, R Shepherd1, A Menzies1, K Leung1, J Teague1, M R Stratton1 and P A Futreal1

1 Cancer Genome Project, Wellcome Trust Sanger Institute, United Kingdom

COSMIC, the Catalogue Of Somatic Mutations In Cancer (www.sanger.ac.uk/cosmic) is designed to store and display somatic mutation information relating to human cancers, combining detailed information on publications, samples and mutation types. The information is curated both from the primary literature and the laboratories at the Cancer Genome Project, Sanger Institute, UK, and then semi-automatically entered into the COSMIC database. The v47 release (May 2010) contained the curation of 9202 papers describing 116,977 mutations across 466,851 samples. In order to provide consistent annotation of the data, COSMIC has developed a classification system for cancer histology and tissue ontology, and adapted HGVS mutation nomenclature recommendations to describe the multiple mutation types involved in cancer.

Cancer genetics is moving from systematic screens of candidate gene sets to whole genome sequencing analyses, and COSMIC displays and navigates this new data; we have recently included systematic gene screens and whole genome sequencing studies. COSMIC will annotate and display somatic mutation data that will be emerging from the International Cancer Genome Consortium (ICGC) and the Cancer Genome Atlas (TCGA) projects. New tools are being developed to interpret this genomic data with coding mutation annotations. In addition COSMIC will be expanded to curate and display data from mouse insertional mutagenesis screening and mouse cancer model exome/genome sequencing in the future. The data within COSMIC is freely available without restriction via a website (www.sanger.ac.uk/cosmic), in datasheets on the FTP site (ftp://ftp.sanger.ac.uk/pub/CGP/cosmic) and through the COSMIC Biomart, available from the COSMIC homepage.

Linking Phenotype Data to the Genome: RGD a Unique Resource for Rat Strains and QTLs

Rajni Nigam, Mary Shimoyama, Mindy Dwinell, Simon Twigger, Diane Munzenmaier, Howard Jacob; the RGD Team

Rat Genome Database, HMGC, Medical College of Wisconsin, Milwaukee, Wisconsin USA

RGD (http://rgd.mcw.edu/) strives to collect all pertinent information related to rat strains and quantitative trait loci (QTLs). At present RGD houses more than 5000 rat, mouse and human QTLs and data for more than 2000 rat strains. QTLs identified in strains give users a new way to maneuver the phenotypes observed and their linkage to the genome. Strains and QTLs contain curated genomic and genetic data which helps researchers in determining specific rat strains that can be used as a disease model for complex human diseases. Our curation is based on a disease-oriented approach where we curate strains that are models for specific human diseases.

We use multiple ontologies to annotate strains and QTLs like Mammalian Phenotype Ontology developed by the Jackson Labs (http://www.informatics.jax.org/searches/MP_form.shtml ) and the Disease Ontology created by MESH at NCBI (http://www.nlm.nih.gov/mesh/MBrowser.html).

Our Phenotypes and Models portal provides another route for physiologists to access genomic and phenotype information. These extend an effort to link phenotype and genotype data which assists users in choosing appropriate model strains for studying phenotypes that closely match those found in human diseases.

RGD has a comprehensive ontology based data structure and annotation system to integrate phenotype data. We use four major ontologies that include the strain ontology, clinical measurement ontology, measurement method ontology and experimental conditions ontology to annotate data from different phenotype projects. These can be used to compare phenotype values for single or multiple strains and researchers can compare their own strains to these, thus linking phenotype data to genomic and genetic variations.
**Logical Operation Based Literature Association with Genes and its application, PosMed.**

Yuko Makita, Rinki Bhatia, Mrinalini Deshpande, Akihiro Matsushima, Manabu Ishii, Yoshiki Mochizuki, Yuko Yoshida, Norio Kobayashi, Tetsuro Toyoda

1Bioinformatics And Systems Engineering (BASE) division, RIKEN., Japan

PosMed prioritizes candidate genes for positional cloning by employing our original database search engine GRASE, which uses an inferential process similar to an artificial neural network comprising documental neurons (or 'documentrons') that represent each document contained in databases such as MEDLINE and OMIM (Yoshida, et al. 2009). PosMed immediately ranks the candidate genes by connecting phenotypic keywords to the genes through connections representing gene–gene interactions other biological relationships, such as metabolite–gene, mutant mouse–gene, drug–gene, disease–gene, and protein–protein interactions, ortholog data, and gene–literature connections.

To make proper relationships between genes and literature, we manually curate queries, which are defined by logical operation rules, against MEDLINE. For example, to detect a set of MEDLINE documents for the AT1G03880 gene in *A. thaliana*, we applied the following logical query: ('AT1G03880' OR 'CRU2' OR 'CRB' OR 'CRUCIFERIN 2' OR 'CRUCIFERIN B') AND ('Arabidopsis') NOT ('chloroplast RNA binding'). Curators refined these queries in mouse, rice and *A. thaliana*. For human and rat genes, we use mouse curation results via ortholog genes in PosMed.

PosMed http://omicspace.riken.jp/PosMed


---

**Epigenetics Knowledgebase –Role Of Epigenetic alterations in Cancer**

Priti Y Ramamohan, Bharat Bhat, Guruprasad Rao, Arathi Raghunath, Uma Rani Periasamy, Lokanath Khamari, Vidyendra Sadanandan, Jignesh Bhave and Jaya Iyer

1Molecular Connections Private Limited, Kandala Mansions, Basavangudi, Bangalore560004, India

Epigenetic genes have emerged as new targets for cancer therapy and the basis for development of new inhibitors and drugs. A repository or a knowledgebase of the epigenome will enable to understand the relationship between neoplasia and epigenetic alterations. At Molecular Connections, we have developed a database, with the aim to bridge the existing knowledge gap between the field of oncology and epigenetics.

The knowledgebase contains data mined from literature (full text articles and patents) for various epigenetic targets of clinical interest. We have manually annotated the molecular details of some of the DNA methyltransferases such as DNMT3A, DNMT3B and histone modification enzymes like EZH2, HDAC6, SIRT1, SIRT2.

The curated data has been organized into following modules: Expression, Inhibitor, Knockout / Knockdown, Mutation, Protein-protein interaction and Metabolite, both in human and animal studies. The information captured is mapped to standard controlled vocabularies (CVs) that includes MeSH, Entrez gene, CAS Ids, SNP as well in-house developed CVs.

The knowledgebase also comes with a user friendly web interface to efficiently query and filter the relevant information from the catalogue of epigenetic genes, for various modules. The distinct records with relevant information can be exported into XML format. The density and the variety of epigenetics related data makes this database a very useful tool in exploring and understanding the process linked with epigenetics.
Clipro™ - A Comprehensive Knowledgebase Of Clinically Significant Proteins In Biofluids

Jaya Iyer1, Vanishree Mallur Srinivas1, Lokanath Khamari1, Vidyendra Sadanandan1, Priti Y Ramamohan1 and Jignesh Bhave1

1 Molecular Connections Private Limited, Kandala Mansions, Basavangudi, Bangalore-560004, India

Quantitative analysis of ‘Biomarker proteins’ in various body fluids for disease diagnosis has attained significant clinical relevance. A curated database of biomarkers in biofluids could aid in analyzing and interpreting the vast pool of available literature data.

We have developed a data repository called Clipro™, a knowledgebase of fluid biomarkers. The knowledgebase contains information manually mined from literature. Clipro™ serves as an effective platform to identify and validate potential biomarker-disease associations through analysis of concentration values in disease versus healthy state. We have systematically compiled a list of biomarkers that contains concentration values in disease states. The potential influence of genetics, sex, age, physiological state and other factors have been captured. The database is compliant with MeSH, GO ontology and drug bank enabling the user to easily integrate and comprehend data.

The pool of data can be navigated through ‘CliPan’, an analysis tool which helps in analyzing the biomarkers- disease associations and predicting prospective targets/biomarkers.

The density and repertoire of information in Clipro™ could be useful for both pharmaceutical companies and researchers in evaluating possible biomarkers for drug discovery.

Challenges to Extract Kinetic Data from Literature

Ulrike Wittig, Renate Kania, Isabel Rojas, Wolfgang Müller

Scientific Databases and Visualization Group, Heidelberg Institute for Theoretical Studies (HITS), Schloss-Wolfsbrunnenweg 35, 69118 Heidelberg, Germany

Biochemical data in the scientific literature is available sparsely in structured and standardised format. For the automatic use of these data databases are developed to provide published and manually extracted data from the literature to the biological community. The grand challenge to develop such databases is the guarantee of the data quality in the database. Therefore the biological experts have to invest a lot of time to read the publications, extract the information, and to insert the data into the database adapting to existing standards. Against the background of the work for the development and curation of the SABIO-RK database (http://sabio.h-its.org) we describe the problems we are confronted with during the process of manual information extraction from the literature and the challenges to ensure a high quality of data in the database. Different examples of publications relevant for the database population will be analysed in detail and problems of missing information and inconsistencies will be presented. Here we separate between issues which could be solved in the future by the development of (semi-)automatic methods to support the biocurators and others which need support from authors or the publisher side.

SABIO-RK is a curated and online accessible database offering comprehensive information about biochemical reactions and their corresponding kinetic data. The database is being populated by merging data from other databases with experimental kinetic data manually extracted from scientific literature or directly submitted from laboratories. The latest version of SABIO-RK now also includes data not only for metabolic but also for signaling reactions.
Manual Curation of Vertebrate Proteins in UniProtKB.

Michael J. Gardner1, Lionel Breuzaad2 and The UniProt Consortium1,2,3

1 European Bioinformatics Institute, Wellcome Trust Genome Campus, Hinxton, Cambridge, CB10 1SD, UK.
2 Swiss Institute of Bioinformatics, Centre Medical Universitaire, 1 rue Michel Servet, 1211 Geneva 4, Switzerland.
3 Protein Information Resource, Georgetown University, 3900 Reservoir Road, Washington, DC 20057-1414, USA.

The UniProt Knowledgebase (UniProtKB) aims to provide the scientific community with a comprehensive, consistent and authoritative resource for protein sequence and functional information. Given the importance of human and vertebrate model data in biomedical research, a major focus is the high-quality manual curation of human proteins and their vertebrate orthologues. Manual curation involves (1) the extraction of experimental results from scientific literature to enrich protein records with a wide range of information including function, structure, interactions and subcellular location, (2) the manual verification of each sequence and clarification of discrepancies between sequence reports, and (3) the assessment of the output of a range of analysis programmes to ensure that sequence features are correctly reported. Manual curation also facilitates the standardization of experimental data – a step necessary for development of methods that enable the semi-automated transfer of manual annotation to uncharacterised or related proteins. Consequently, manual curation of vertebrate proteins plays a vital role in providing users with a complete overview of available data while ensuring its accuracy, reliability and accessibility. UniProtKB/Swiss-Prot currently contains the complete manually reviewed human proteome, comprising approximately 20'300 proteins, and an additional 61'000 reviewed entries from model vertebrates such as mouse, rat, apes, cow, chicken, zebrafish and Xenopus. Ongoing efforts continue to improve the quality of vertebrate sequences in collaboration with HAVANA, Ensembl, HGNC and RefSeq, to include new functional information as it becomes available, and to extend the coverage of curated proteins in vertebrate species. All data are freely available from www.uniprot.org.

Manual Biocuration, A UniProtKB/Swiss-Prot Perspective

Sylvain Poux1, Ioannis Xenarios1,2 and the UniProt consortium1,3,4

1 Swiss Institute of Bioinformatics, Switzerland
2 Vital-IT Group, Switzerland
3 European Bioinformatics Institute (EBI), UK
4 Protein Information Resource (PIR), USA

Manual curation of proteins is essential to provide high quality datasets that can then be programmatically parsed and allow the scientific community a rapid access to reliable information. Manually curated entries are stored in the UniProtKB/Swiss-Prot section of UniProtKB and constitute the highest priority of the UniProt consortium, with more than 60% of its staff being fully dedicated to this task. High quality information is added by experienced biocurators, most of them having a strong background in research. Priority is given to the biocuration of proteins with impact on the largest number of users. Biocurators extract a maximum of pertinent information from the scientific literature. Information is reviewed, compiled, summarized and reported in the appropriate fields of a UniProtKB entry. Controlled vocabularies are used to structure the information of an entry and evidence tags are attached to each piece of information indicating its source. We also assign Gene Ontology (GO) terms to all UniProtKB entries during the biocuration process, by extracting information related to each of the GO ontologies with experimental evidence.

To avoid duplication of annotation efforts with other resources we establish collaborations with a number of databases (i.e. MODs) to ensure an accurate and mutually beneficial integration pipeline. Collaborations have started in fields such as proteomics and protein-protein interactions and future plans involve collaboration with the NCBI’s Genome project and the Reference Sequence (RefSeq). This constant effort and the experience accumulated over the years has created a unique task force in the domain of manual biocuration.
Gene Ontology Annotation of Hearing

Guvanch Ovezmyradov¹, Martin Goepfert¹

¹Department of Cellular Neurobiology, University of Goettingen, Max-Planck-Institute for Experimental Medicine, Hermann-Rein-Str. 3, 37075 Goettingen, Germany

Hearing impairment is a common sensory disease. Recent progress in identifying and characterizing deafness genes lead to a better understanding of the molecular processes involved. Together with the experts in the field, we are annotating the auditory-relevant genes in humans and model organisms using the Gene Ontology annotation best practices. The Auditory Gene Ontology Initiative is aimed to capture information for both humans and model organisms, including Drosophila, zebrafish and mouse. The annotations will be uploaded to UniProtKB-GOA, FlyBase, ZFIN and MGI. This will facilitate cross-species comparisons of genes for hearing. Based on the annotations, a systematic analysis is presented that provides an overview of the molecular functions and cellular localizations of auditory genes.

caNanoLab – A Tool To Benefit Biomedical Nanomaterials Research

Michal Lijowski¹, Nathan Baker²

¹ Washington University in St. Louis, USA
² Pacific Northwest National Laboratory, USA

The cancer Nanotechnology Laboratory Web portal (http://cananolab.nci.nih.gov/caNanoLab/welcome.do) is designed to promote sharing of nanomedicine data and to provide support for the annotation of nanoparticles with information about their composition, function, and characterization. The project is a part of NCI cancer Biomedical Informatics Grid (caBIG®), and includes a collaboration among the NCI Center for Biomedical Informatics and Information Technology (CBIIT), the NCI Nanotechnology Characterization Laboratory (NCL), and the NCI Alliance for Nanotechnology.

Currently, caNanolab contains nanomaterial data of 765 "public" samples. 18 samples were contributed by National Characterization Lab (USA), while the remaining 747 were submitted by a curator at Washington University in St. Louis from 114 publications. The curated samples include simple nanoparticles, such as emulsions, metal oxides, nanohorns, nanorods, polymers, quantum dots, as well as complex particles, such as dendriworms, which are conjugates of crosslinked dextran-coated iron oxide particles and dendrimers.

Data submission involves curation of sample composing elements, their function (imaging, targeting, therapeutic), physico-chemical (size, molecular weight, surface/zeta potential, etc.) properties, and in vitro (cytotoxicity, blood contact, targeting/gene expression, etc.) characterizations, based on information provided in corresponding publications or obtained from the authors.

caNanoLab also contributes to and utilizes resources of NanoParticle Ontology (http://www.nano-ontology.org).
ACBD: Database for Ascidian Chemical Genomics

Yuichiro Hira1, Jun Terai1, Mitsuru Nakamura1, Etsu Tashiro1, Masaya Imoto1, Kotaro Oka1, Kohji Hotta1

1Dept. of Biosci. and Informatics, Factulty of Sci. And Tech., Keio Univ., Japan

Chemical biology approach enables us to understand complex biological systems, using small molecules such as specific inhibitors or activators of proteins, hormone-like inducers, and neurotransmitters. When the analysis is done genome-widely, that research is called chemical genomics. We are now planning to make a new start of chemical genomics using ascidian, Ciona intestinalis. Because Ciona have many advantages for screening mutant phenotypes. "high-throughput": relatively shorter developmental time (hatching after 18hr) than other model organisms. "high-resolution": the size of ascidian larva is relatively small (~2mm) and transparent so a confocal microscope can image a whole-animal at a single cell level.

As a first step for starting ascidian chemical genomics, we are constructing a new database called ACBD (Ascidian Chemical Biology Database). ACBD has two main parts. One is “Already-used chemical” part. First of all, we searched chemicals that are used at ascidian from around 800 past ascidian articles searched from pubmed. Then it was found that 278 kinds of chemicals have been used already in ascidians, these chemicals are curated by each effect in ascidian. The other is “Not-yet-used chemical part.” This part is mainly based on the information from DrugBank (Wishart DS et al., 2008). From DrugBank, we got 4,537 protein sequences of drug targets protein. After that, we did blast search by using that protein sequences and all Ciona KH models (24,000 models) as queries. As a result, we found that 2,778 KH models could be the targets of chemicals. In near future, ACBD will be open on the web.

Towards Automated Annotation Of Ortholog Clusters Using Natural Language Processing Technologies

Yasunori Yamamoto1, Shinobu Okamoto1,2, Mitsuteru Nakao1,2, Takatomo Fujisawa2, Yasukazu Nakamura2,3, Toshihisa Takagi1,3

1Database Center For Life Science, Japan
2Kazusa DNA Research Institute, Japan
3DNA Data Bank of Japan, Japan

A large number of ortholog databases have been developed. They not only decide ortholog clusters (OCs) from various sources to allow comparison between spices but also aim to provide annotation tools as sets of reference sequences. The practical usability of OCs is determined by the ability to provide an explicit functional annotations. However, studying a protein function based on experiments takes much longer time than obtaining a protein sequence, and many proteins are not based on experiments but on computational analysis such as those beginning with similar to or probable. Therefore, those less credible protein annotations frequently appear in OCs. In this study, we use Natural Language Processing (NLP) technologies to extract the most reliable protein annotation from those in a cluster.

We first surveyed how protein annotations are composed using the eggNOG database, then obtained term frequency distributions and frequently appearing patterns. For each pattern, we classify it to a originally defined class based on its reliability from a curator's point of view. Next, we try to obtain a model of protein annotation structures from various sources using a machine learning approach that is well studied in the field of NLP. This can be used to rank a given protein annotations in an OC, which are shown at the poster.
Information on Transcriptional Regulation and Signal Transduction of *Escherichia coli* K-12 Integrated in the Database RegulonDB.

Programa de Genómica Computacional, Centro de Ciencias Genómicas, Universidad Nacional Autónoma de México, A.P. 565-A, Cuernavaca, Morelos 62100, México.

Since its inception, RegulonDB (http://regulondb.ccg.unam.mx/) has been a database that compiles information about the regulation of transcription initiation of *Escherichia coli* K-12. However, we are aware that transcriptional regulation is not an isolated process; instead, it is the response to the different environmental conditions that trigger a series of concatenated reactions that end in transcriptional regulation, and it implies an adequate response in terms of induced and repressed gene products. We are working now to include all these new data in RegulonDB. As a consequence, transcriptional regulation in RegulonDB will be part of a unit that initiates with the signal, continues with the signal transduction to the core of regulation to modify expression of the affected set of target genes, and ends with an adequate response. We refer to these units as genetic sensory response units, or geSorgans.

The inclusion of geSorgans will bring a dramatic change and expansion of RegulonDB, due to the fact that we will be adding several new types of reactions and interactions. We started to collect data about signal transduction of the sigma factors, the two-component systems, of some transcription factors involved in carbon source utilization, and of genes involved in the synthesis of amino acids. We plan a high-level curation with super-pathways summarizing concatenated sets of reactions linked to those other databases that curate such information, while enabling with RegulonDB a compilation of complete geSorgans.

In addition, the number of DNA binding sites for some transcription factors has grown considerably, and therefore we decided to review systematically those sites whose lengths range from 40 to 60 bp with orientation and consensus sequences that are not easy to identify. The current version of RegulonDB is the beginning of a higher-level curation of gene regulation information, and eventually our database will include all regulatory mechanisms and their regulated genes.

The BioGRID Interaction Database

Livstone M\(^1\), Breitkreutz BJ\(^2\), Stark C\(^2\), Boucher L\(^2\), Chatr-Aryamontri A\(^3\), Oughtred R\(^1\), Nixon J\(^3\), Reguly T\(^2\), Rust J\(^1\), Winter A\(^3\), Dolinski K\(^1\), Tyers M\(^2\)\(^3\).

(1) Lewis-Sigler Institute for Integrative Genomics, Princeton University, Princeton, NJ, USA; (2) Samuel Lunenfeld Research Institute, Mount Sinai Hospital, Toronto, Canada; (3) Wellcome Trust Centre for Cell Biology, The University of Edinburgh, Edinburgh, UK.

The goal of the Biological General Repository for Interaction Datasets (BioGRID) (http://www.thebiogrid.org) is to archive and freely disseminate collections of genetic and protein interactions from major model organisms. BioGRID currently houses over 335,000 interactions curated from high-throughput datasets and individual focused studies found in the primary literature, as derived from some 23,000 publications. Complete coverage of the entire literature for both the budding yeast *Saccharomyces cerevisiae* and the fission yeast *Schizosaccharomyces pombe* has been achieved, resulting in the curation of over 246,000 interactions, and efforts to expand curation across multiple species are underway. Through collaborations with the Gene Ontology (GO) Consortium and the Linking Animal Models to Human Disease Initiative (LAMHD), we are focusing our curation efforts across model organisms on particular areas of biology to enable insights into conserved networks and pathways that are relevant to human health.

The BioGRID 3.0 web interface contains new search and display features that enable rapid queries across multiple data types and sources. A dedicated Interaction Management System (IMS) is used to track all curation and to prioritize publications across multiple curation projects. BioGRID data are incorporated in several model organism databases and other biological databases. The entire BioGRID interaction collection may be downloaded in multiple file formats, including PSI MI XML, and source code for BioGRID is freely available without any restrictions. This work is supported by NIH NCRR grant R01 RR024031 to MT and KD, and by grants from the CIHR and BBSRC to MT.
Ontology on Cell Lines

Arathi Raghunath, Bharath Bhat, Jignesh Bhate and Usha Mahadevan

Molecular Connections, Pvt. Ltd., Kandala Mansions, 2/2, Kariappa road, Basavanagudi, Bangalore, India

Ontologies provide a common platform of controlled vocabulary for researchers who need to share information across a particular domain, inclusive of machine-interpretable definitions of basic concepts in the domain and relations among them. The need for ontologies that provide a systematic arrangement of available data and allow cross talk across various related domains is gaining momentum. In this backdrop, we have developed a comprehensive ontology on primary and established cell lines—both normal and pathologic. It covers around 1500 cell lines. This ontology has been built to include the major domains in the field of biology like anatomy, bio-molecules, chemicals and drugs, pathological conditions and genetic variations around the cell lines. An extensive network of relations has been built across these concepts to enable different combinations of queries. The ontology is built in OWL format, in compliance with the OBO foundry standards.

The ontology covers all cell lines from major sources like ATCC, DSMZ, ECACC, ICLC etc. All details available for a given cell line are captured in the ontology. Synonyms for the cell lines and the catalogue numbers from the above mentioned sources are included in the ontology—allowing query for cell lines using the catalog numbers or synonyms. Synonyms for entities other than cell lines are also covered. A brief description for the entities is provided as annotations. The ontology includes terms from standard public domains like MeSH, NCBI etc., wherever possible. Application of this ontology in building other ontologies like bioassay ontology is discussed.

eagle-i: An Ontology-Driven Framework For Biomedical Resource Curation And Discovery

Erik Segerdell1,3, Melanie L. Wilson1,3, Ted Bashor2,3, Daniela Bourges-Waldegg2,3, Karen Corday2,3, H. Robert Frost2,3, Tenille Johnson2,3, Larry Stone2,3, Carlo Torniai1,3 and Melissa A. Haendel1,3

1Oregon Health & Science University, USA
2Harvard University, USA
3The eagle-i Consortium, USA

The eagle-i Consortium (www.eagle-i.org/home) comprises nine geographically and ethnically diverse universities across America working to build a federated network of research resources. Biomedical research generates many resources that are rarely shared or published, including: reagents, protocols, instruments, expertise, organisms, training opportunities, software, human studies, and biological specimens. The goal of eagle-i is to improve biomedical research by helping researchers more easily find scientific resources that are difficult to discover, reducing time-consuming and expensive duplication of resources. Now in early development, the system will ultimately expand to include research resources at other universities following the end of the two-year pilot phase. An application ontology is being developed to enable representation of core facility and research lab resources in the eagle-i repository, leading to more effective searches and better linkage between data types. The eagle-i ontology will guide users to valid queries via auto-suggestion, ontology browsing, concept-based search, and synonym expansion. The ontology development effort is being guided by active discussions within the ontology community and brings together relevant preexisting ontologies in a logical framework. Components of the data entry and search interfaces are generated directly from the ontology, which allows rapid change in response to user needs and ontology evolution. Each eagle-i institution will populate and manage a local repository using data collection and curation tools. To enhance the quantity and quality of data, the data tools will take advantage of the ontology to support semi-automated annotation of resources. NIH/NCRR ARRA award #U24RR029825.
Using Multiple Ontologies to Curate Phenotype Data

Mary Shimoyama, Rajni Nigam, Melinda Dwinell

Rat Genome Database, Medical College of Wisconsin, United States

The strength of the rat as a model organism lies in its utility in pharmacology, biochemistry, and physiology research. Data resulting from such studies is difficult to represent in databases and the creation of user-friendly data mining tools has proven difficult. The Rat Genome Database has developed a comprehensive ontology-based data structure and annotation system to integrate phenotype data along data on samples, measurement methods and experimental conditions. RGD uses standard data formats and multiple ontologies to integrate phenotype data from large scale phenotype projects, QTL studies and other projects. The four major ontologies include the strain ontology, clinical measurement ontology, measurement method ontology and experimental conditions ontology. This annotation system has facilitated curation of phenotype data from the literature as well as from large scale phenotyping projects. In addition, the multiple ontology approach has provided an excellent platform for a dynamic data mining and visualization tool which allows the user to search for phenotype information using a series of filters based on each of the ontologies. Search results are visualized in charts and are available in download files for easy comparisons of phenotype readings for single or multiple strains, conditions or measurement methods. A further aspect of this project is the development of curation software for greater efficiency and accuracy in the curation process. Continued developments will include data submission software to allow researchers to integrate their own data at RGD and data analysis tools.

Disease Ontology: Logical Definitions and Relations

LaRon Hughes1, Suvarna Nadendla2, Lynn Schriml2, Warren Kibbe1

1NU Clinical and Translational Sciences Institute (NUCATS), Northwestern University, USA
2Center for Bioinformatics and Computational Biology, University of Maryland, USA

The Disease Ontology (DO) is a community driven, open source ontology that is designed to link disparate datasets through disease concepts for the integration of biomedical data associated with human disease. DO provides a semantically computable structure of inherited, environmental, and infectious human disease to connect phenotype, symptom, genetic, microbial and clinical data through the lens of human disease. One of the scopes of the DO project is to create logical definitions, which complements the textual definitions to allow computer agents to make sense of the data (e.g., cross-ontology references). To facilitate logical definitions, several relations have been created such as results in, transmitted by, and has agent that are relevant to DO. The combination of robust textual definitions, logical definitions, and relations facilitates a more useful resource.
Microbial Genome Annotation at NITE and Annotator-Oriented Development of Annotation Workbench

Natsuko Ichikawa¹, Toshihiko Sawano¹, Tomohiro Sekigawa¹, Hidekazu Nakazawa¹, Shoko Ohji¹, Shuji Yamazaki¹ and Nobuyuki Fujita¹

¹Department of Biotechnology, National Institute of Technology and Evaluation (NITE), Japan

NITE have been engaged in whole genome analysis of microorganisms each representing taxonomically and industrially important groups since 1998. To date, we contributed to over 40 genome projects and released 20 genomes in collaboration with universities, public institutions and private sectors. We also continue to provide accurate manual annotations including re-annotation of published genome such as Aeropyrum pernix. We have been maintaining an original annotation workbench called OCSS since 2001, which is flexible for (i) sequence update, and suitable for (ii) reference collection and (iii) group annotation in annotator-friendly manner.

We will present our annotation methods and implemented functions of OCSS. Taking advantages of in-house development, many of the functions have been implemented to OCSS based on annotators’ request.

To help early start of annotation for unfinished contigs, the annotation data are maintained at the ORF level, so that annotation can be easily allocated to finished sequences. Various analyses such as BLASTP and InterProscan are automatically executed when ORF is newly assigned or updated. We adopt literature information as supporting evidences of annotation. The OCSS automatically provides literature information of known homologues together with the list of homologous proteins generated by BLASTP analysis. The assignment records of literature are shared among genome projects and annotators, for facilitation of annotation speeds. From the very first of its development, the OCSS system was intended for use in group annotation. The history management and job assignment systems were implemented to follow up each annotator’s contribution and avoid unnecessary double commitments.

Consequently, OCSS is a suitable workbench for primary annotation in sequencing centers.

ENA as an Information Hub

Petra ten Hoopen, Ruth Akhtar, Richard Gibson, Bob Vaughan, Nadeem Faruque and Guy Cochrane

ENA, EMBL-EBI, Wellcome Trust Genome Campus, Hinxton, Cambridge, United Kingdom

The European Nucleotide Archive (ENA; http://www.ebi.ac.uk/ena/) is a comprehensive repository for public nucleotide sequence data from nearly four hundred thousand taxonomic nodes. Together with partners in the International Nucleotide Sequence Database Collaboration (INSDC; EBI, NCBI and DDBJ) we provide a broad spectrum of sequences, from raw reads (Sequence Read Archive data class), assembled contigs (Whole Genome Shotgun data class), assemblies of EST transcripts (Transcriptome Shotgun Assembly data set), to partial or complete assembled nucleic acid molecules with functional annotation derived from direct and third party experimental evidence (Standard and TPA data classes, respectively).

Resources beyond ENA, such as RNA and protein databases, genome collections and model organism services, use data stored and presented at ENA as both source and underlying supporting evidence for their records. Integration of the growing wealth of molecular information is a great challenge that brings opportunities for ENA to serve as a bioinformatics data information hub, allowing, through its provision of permanent identifiers for sequence and project records, community-recognized identifiers for navigation across databases.

As a comprehensive repository of directly sequenced nucleic acid molecules we have the unique opportunity to obtain exact provenance information directly from the submitting researchers. Our pre-publication biocuration efforts are focused on obtaining rich and accurate information on the sample that has been sequenced and on the methodology surrounding its preparation for sequencing. We present here an insight into data flow in the archive and a straightforward biologist-orientated submission system with a rule-based validator for smaller sets of sequences.
Gene Trek in Prokaryote Space (GTPS) reduces GIGO in the annotation of microbial genome sequences

Takehide Kosuge1, Yaumasa Shigemoto2, Kuwana Yoshikazu3, Hideaki Sugawara1

1National Institute of Genetics, Japan
2Fujitsu, Japan
3Tokai software, Japan

More than 1,200 microbial genome sequences and annotations are publicly available from the International Nucleotide Sequence Databases (INSD) of DDBJ/EMBL/GenBank. However, the annotations released from INSD are often carried out by the different protocols, including minimum length of the ORF in the ORF prediction, threshold value of BLAST search and version number of the reference data used for BLAST and motif scan. In addition, it is highly probable that erroneous ORFs are propagated to the annotation of newly sequenced microbial genomes. The purpose of GTPS (http://gtps.ddbj.nig.ac.jp/) is to re-annotate the ORFs among microorganisms in INSD by using a common protocol and diffuse the results to every user as a resource for genome scale analysis on microbes.

In GTPS, the ORFs are classified into from AAAA (top grade) to X (lowest grade) categories by using the result of BLAST and InterProScan analysis. We provide you with all the result of re-annotated data by the graphical user interface and the flat file. GTPS database has been updated every year since 2003. The number of re-annotated strains/ORFs increased from 123/283,209 to 862/2,838,183 in 2003-2009. The ORFs in 2009 are distributed in each grade as follows: 2,340,865 in AAAA-A, 26,900 in BBBB-B, 39,338 in C, 431,270 in D, 53,106 in E and 1,173,400 in X. GTPS in 2009 newly predicted 93,045 ORFs. This is a significant number, although the ratio of the new ORFs to ORFs in INSD was 3.3%. The statistics in details and the protocols will be presented at the conference.

GTPS is supported by Institute for Bioinformatics Research and Development, JST.

Automatic protein clustering as a basis of automatic annotation

Naoki Sato1

1Graduate School of Arts and Sciences, University of Tokyo, Japan

Development of new generation sequencers enabled genome sequencing feasible for every organism in a laboratory. A typical data flow of de novo sequencing includes (1) assembly of sequence reads, (2) estimation of open reading frames, (3) annotation of proteins, and (4) finding RNA genes. The annotation is normally performed by BLASTP searches against several different databases. However, it is usually hard to find a plausible annotation by just looking at the results of BLAST searches.

Here I propose a potentially automatic method of annotation that exploits automatic protein clustering using the software GCLUST, which estimates proper similarity threshold for each list of homologs using ‘entropy-optimized organism count’ method (Sato 2009). The software has been used to construct a homolog database including both prokaryotic and eukaryotic proteins (http://gclust.c.u-tokyo.ac.jp/). For use in genome annotation, we need de novo clustering including many genomes of related organisms as well as genomes of representative organisms. Application of protein clustering in the annotation in Arthospira platensis was the first successful case (Narikawa et al. 2010). I present here results of protein clustering of total predicted proteins in two draft genomes of cyanobacteria along with total predicted proteins of 41 cyanobacteria available at NCBI. For each of the resultant protein clusters, an alignment and a phylogenetic tree were also prepared for assistance in functional annotation. The quality of alignments was evaluated by counting ill-aligned proteins (missing N- or C-terminus, or insertion/deletion), which was 4-13% of total predicted proteins in most cyanobacterial genomes. Annotation may be automated by extracting significant key words already assigned for member proteins of clusters or by comparison with reference protein clusters.
The Challenge Of Eukaryote Genome Annotation At Genoscope

Betina M. Porcel, Franck Aniere, Sylvain Bonneval, Benjamin Noel, Jean-Marc Aury, Corinne Da Silva, Olivier Jaillon, France Denoeud, Claude Scarpelli, Jean Weissenbach, Patrick Wincker and François Artiguenave.

Genoscope, Institut de Génomique, CEA, 2 Rue Gaston Crémieux, CP 5706, 91057 EVRY cedex, France

Annotating a genome aims at assigning valuable biological information on genomic sequences. The annotation process includes different steps, starting by the definition of the gene structures and ending by the functional annotation of protein functions and the definition of homology relationships. Finding these genes involves computational methods as well as experimental validation.

As a consequence of the adaptation of the novel generation sequencing machines, the number of eukaryote genomes sequenced and annotated at Genoscope is increasing constantly. For the last 5 years, we have been annotating over 10 eukaryote genomes sequenced and annotated at Genoscope is increasing constantly. For the last 5 years, we have been annotating over 10 eukaryote genomes using our Eukaryote Genome Annotation workflow. We will present few examples of annotated genomes as well as our developments to automate the annotation process which allow us to annotate today 2 to 4 genomes a year.

However, even though automatic computational predictors are extremely useful for a large scale annotation at a preliminary level, their predictions are sometimes inadequate. Therefore, any gene models require human expertise to find errors and resolve incongruous evidence on the automatic annotation of the genome. In order to achieve both high-throughput and high quality of annotations, we set up a scalable model for Community Genome Annotation, which combines automated annotation, community-wide genome analysis and manual validation based on GMOD components (Chado/GBrowse/Apollo). We provide to our collaborators a distributed annotation platform allowing an expert evaluation of genome annotations, in addition to our automated gene prediction pipeline.

Training biocurators in annotation jamborees, as well as coordination and support for the user communities have been launched, as it will be shown. This community based-manual annotation facilitates expert networking and comparison of predicted annotations with existing biological data.

ANNOTATION STRATEGIES FOR HIGH-COVERAGE, LOW-COVERAGE AND NEXT GENERATION SEQUENCING ASSEMBLIES IN ENSEMBL

Vogel J(1), Aken B(1), Fairley S(1), Hourlier T(1), Ruffier M(1), Tang A(1), White S(1), Zadissa A(1), Hubbard T(1), Searle S(1)

(1) Wellcome Trust Sanger Institute

Ensembl is one of the leading sources of automatic genome sequence annotation data for a variety of different genomes. Ensembl annotation is currently available for more than 50 different genomes, ranging from high-quality reference assemblies like human or mouse over low-coverage assemblies to short-read assemblies, which are based on next-generation sequencing technologies.

To cope with such a variety of different assembly characteristics, we have developed different genome annotation strategies to suit the specific requirements of each genome. For reference genomes like Human, we combine the available manual genome annotation generated by the HAVANA group with our automatic annotation data to construct a complete gene set - the GENCODE set. For low-coverage species, we developed a system to project genome annotation of an evolutionary close, well-annotated reference species through genomic alignments onto the low-coverage assembly of the target species. This approach can also be used to annotate genomes whose assemblies are based on short-read sequencing.

With the recent developments of next-generation sequencing technologies and the availability of transcriptome sequence data, we have now started to incorporate this new data types into our genome annotation pipelines, starting with the zebrafish genome annotation for the ZV9 assembly – this annotation will be released in late 2010. Other evidence sources used for genome annotation comprise ChipSeq data, comparative data and cDNA alignments – for example for the identification and annotation of long non-coding genes. The annotation for all Ensembl-annotated genomes can be viewed at www.ensembl.org, downloaded from ftp.ensembl.org, and accessed via various APIs, the data-mining tool BioMart and MySQL.
Extension of Database of Transcription Start Sites (DBTSS) with a Large Number of Sequences and Analysis of Bi-directional Promoters

Riu Yamashita1,2, Hiroyuki Wakaguri3, Yutaka Suzuki3, Sumio Sugano3, Kenta Nakai2

1 Frontier Research Initiative, Institute of Medical Science, University of Tokyo, Japan
2 Human Genome Center, Institute of Medical Science, University of Tokyo, Japan
3 Department of Medical Genome Science, University of Tokyo, Japan

The annotation for a huge number of sequences from next generation sequencers is one of the indispensible steps of biological data analysis. We have constructed the DataBase of Transcription Start Sites (DBTSS: http://dbtss.hgc.jp/) to support the analysis of transcription regulation research. DBTSS contains the information of accurate transcription start sites (TSSs) based on experimentally determined 5’-end clones. In last year, we updated the database, and now DBTSS has 328 million new short sequences generated by massively sequencing the 5’ end of oligo-cap selected cDNAs in humans and mice. These short sequences were collected from normal fetal or adult human tissues, including brain, thymus, liver, kidney and heart, from 6 human cell lines in 21 diverse growth conditions as well as from mouse NIH3T3 cell line: altogether 31 different cell types or culture conditions are represented.

Using DBTSS data, we have observed several biological new aspects. Here, we focused on bi-directional promoters, which has TSSs on both plus and minus strand closely. Among 177,125 TSSs, the distribution of the distance between plus and minus strand TSSs showed significant two peaks. The first peak corresponded to less than 100bp (7,445 TSSs), and the second corresponded to one over 1000 bp (70,194 TSSs). We also observed significant characteristics difference in the first class of bi-directional promoters: namely, more tissue-specific expression, highly ordered nucleosome structures. These results indicate that our database provides a powerful solution not only to discriminate TSSs having clear biological significance from the other possible noise level transcriptions, but also to analyze transcription regulation. Also, these data provides the dynamical change of transcription start sites in different cell types and conditions.

The Hymenoptera Genome Database

Monica C. Munoz-Torres1, Justin T. Reese1, Christopher P. Childers1, Anna K. Bennett1, Jaideep P. Sundaram1, Donald C. Vile1, Christine G. Elsik1.

1Georgetown University, Department of Biology, Washington, DC 20057. USA.

The Hymenoptera Genome Database (HGD) is an informatics resource supporting genomics of hymenopteran insect species. This relational database implements open-source software and components providing access to curated data contributed by an extensive, active research community. HGD includes the genome sequences and annotation data of honey bee, Apis mellifera (http://BeeBase.org) the parasitoid wasp, Nasonia vitripennis (http://NasoniaBase.org) and a portal to the genomes of six species of ants. Together, these species cover approximately 200 MY in the phylogeny of Hymenoptera, allowing to leverage genetic, genome sequence, and gene expression data, as well as the biological knowledge of related model organisms. The availability of resources across an order greatly facilitates comparative genomics and enhances our understanding of the biology of agriculturally important Hymenoptera species through genomics. HGD has orchestrated research contributions from an extensive community from almost 70 institutions in 13 countries, constituting what is perhaps the largest dispersed manual annotation effort reported to date. Community annotation efforts are made possible thanks to a remote connection to a Chado database by Apollo Genome Annotation client software. Curated data at HGD includes predicted and annotated gene sets supported with evidence tracks such as ESTs/cDNAs, small RNA sequences and GC composition domains. Data at HGD can be queried using genome browsers and / or BLAST/PSI-BLAST servers, and it may also be downloaded to perform local searches. We encourage the public to access and contribute data to HGD at http://HymenopteraGenome.org.
A Trial For Genome Re-annotation Using the Literature and Annotation Reference Information

Takatomo Fujisawa1, Shinobu Okamoto1,2, Mitsuteru Nakao1,2, Yasukazu Nakamura1,3
1Kazusa DNA Research Institute, Japan
2Database Center For Life Science, Japan
3DNA Data Bank of Japan, Japan

The next-generation sequencers have significantly accelerated whole-genome sequencing and re-sequencing. However, for the functional annotation, it remains a hard task. Therefore, it is important to give information with high efficiency and quality to the genome features. For the automation of high-quality annotation, it need to improve the accuracy of each step to 1) classify by clustering and/or mapping method, 2) assign general product name and symbol name, and 3) assign these for species-specific annotation.

We have accumulated the literature annotations that are manually extracted the gene/product name and literature section information from full papers in KazusaAnnotation (http://a.kazusa.or.jp/). It is designated as the Gene Indexing (GI), which is named-entity recognition for gene molecule. We have re-annotated on the plant-related microorganism genomes by using GI data of the gene/protein symbol and Pubmed ID and these literature Information. To classify the inferred functional annotation of information is possible by using the publication data on almost all the genes. We believe that it is useful and valuable to update reference genome annotations using the latest literature information in gene/protein names for species-specific annotation.

We will report and discuss about the functional annotation with the reference information stored in the genome annotation jamboree.
Analyzing Inverse Symmetry with Original and Terminal Sites of Prokaryotic Genomes

Tzu-Ting Hsu¹, Hoong-Chien Lee¹

¹Graduate Institute of Systems Biology and Bioinformatics, National Central University, Chungli, Taiwan, Republic of China

The segmental duplication has long been known to be a major mechanism for genome growth and evolution. Beyond that, recent study has spotted occurrence of whole-genome duplication on yeast. In this work we investigate possible association of inverse symmetry with whole-genome inverse duplications on genomes. Our findings, through analyses of word-frequency nucleotides and distributions of homologous conserved regions on publicly available complete genomes of 18 archaea and 139 bacteria, are positive. These findings suggest, first, that whole-genome inverse duplications have occurred in most prokaryotic chromosomes near original sites and terminal sites. Secondly, inverse symmetry can be taken as a feature to predict loci of original sites and terminal sites in prokaryotic chromosomes. Our research integrates the abundance of knowledge with evolution of genomes and creates a new approach for predicting loci of original sites and terminal sites in prokaryotic chromosomes.

An Italian SOLanaceae Integrated Platform (ISOL@): the Kick Start to exploit Solanaceae Comparative Genomics in Solanaceae

Alessandra Traini¹, Nunzio D'Agostino¹, Miriam Di Filippo¹, Domenico Carputo¹, Luigi Frusciante¹ and Maria Luisa Chiusano¹

¹Dept. of Soil, Plant, Environmental and Animal Production Sciences, University of Naples Federico II

The long-term goal of the International Solanaceae Genome Project is to exploit the information generated from ‘omics’ on Solanaceae species, in order to analyze the genome organization, the functionality and the molecular evolution of the entire Solanaceae family. This encouraged the production of an overwhelming amount of molecular data requiring advanced computational technologies to properly offer bioinformatics resources for suitable investigation and data mining. In this frame, we designed and implemented the multilevel computational environment ISOL@, an Italian SOLanaceae resource (Chiusano et al., 2008). In its early life, ISOL@ was exclusively focused on the analysis of the BAC-based tomato genome sequence, but it was constantly updated and evolved to adapt to novel technologies and to include other Solanaceae species. At present, ISOL@ includes tomato (Solanum lycopersicum) and potato (Solanum tuberosum) genome sequences and the publically available Solanaceae transcriptomes, collected in the SolEST database (D’Agostino et al., 2009), offering accessory applications to improve cross referencing among data and profiting from the integration of various, heterogeneous collections from different species. The gathering and convergence of data generated by high-throughput technologies, the effective integration of different collections and the analysis of the information content based on comparative approaches represent the challenges that ISOL@ attempts to solve. Indeed, we believe they represent key approaches to support ‘-omics’ efforts and meaningful interpretations to exploit and understand biological relevant information from intriguing data sources.

Work supported by the Agronanotech (Ministry of Agriculture, Italy), the GenoPom (Ministry of University and Research, Italy) and the EUSOL Projects (EU VI Frame Programme).
H-DBAS: Human-transcriptome DataBase for Alternative Splicing, Version 6

Jun-ichi Takeda\textsuperscript{1,2}, Yutaka Suzuki\textsuperscript{2}, Sumio Sugano\textsuperscript{2}, Takashi Gojobori\textsuperscript{1,3} and Tadashi Imanishi\textsuperscript{1}

\textsuperscript{1}Integrated Database and Systems Biology Team, Biomedicinal Information Research Center, National Institute of Advanced Industrial Science and Technology, Japan
\textsuperscript{2}Department of Medical Genome Sciences, Graduate School of Frontier Sciences, the University of Tokyo, Japan
\textsuperscript{3}Center for Information Biology and DDBJ, National Institute of Genetics, Japan

H-DBAS is a unique database of human alternative splicing (AS). In this database, representative AS variants (RASVs) are identified for each locus using comprehensive datasets of human full-length cDNA and mRNA sequences in H-InvDB. Because we used sequences of curated transcripts, the resulting annotation of human RASVs and other feature are highly reliable. H-DBAS was made public in 2006 for the first time to present the human RASVs affecting protein function by genome-wide analysis. Since 2006, H-DBAS was updated annually by adding new annotation results. For example, we added the conserved AS between human and mouse identified by using comparative genomics, the AS junctions specific to a certain cellular fraction by using RNA-Seq results, and so on. Recently, as a latest version 6, we refurbished the total design of H-DBAS and increased the annotation data considerably. In this major update, in order to support functional prediction of AS variants based on the evolutionary conservation, human RASVs were compared with those of five model mammalian organisms such as mouse, rat, chimpanzee, macaque and dog by using alignments of genome and transcript sequences. Consequently, we identified equally-spliced variants (ESVs) that corresponding to conserved splicings between human and above species, the numbers being 5071, 2717, 7679, 7116 and 4261, respectively. In addition to the comparative transcriptomics analysis among mammalian species, H-DBAS included the information of splice enhancer and silencer motifs as well as the positions of human SNPs. These annotations are displayed in a reconstructed and improved viewer of H-DBAS. H-DBAS Version 6 is available at http://www.h-invitational.jp/h-dbas/.

Cancelled
Evolutionary Analysis of Nucleotide Sequences Using G-compass, a Web-based Comparative Genome Browser between Human and Vertebrates

Yuichiro Hara1, Ryuichi Sakate1,2, Takashi Gojobori1,3, and Tadashi Imanishi1

1Biomedicinal Information Research Center, Advanced Industrial Science and Technology, Japan, 2National Institute of Biomedical Innovation, Japan, 3Center for Information Biology and DNA Data Bank of Japan, National Institute of Genetics, Japan

G-compass (http://hinv.jp/g-compass/) is a web-based genome browser providing one-to-one genome alignments between human and 12 vertebrates. G-compass provides three ways of alignment visualization, comparative genome browser, alignment sequence viewer, and dot plot viewer, which easily enable us to detect synteny conservations and genomic rearrangements on the web browser. In addition, the aligned regions are linked to not only the curated functional annotations of the human genes and transcripts from H-InvDB (http://www.hinv.jp/hinv/ahg-db/) but also the genome feature tracks that are annotated uniquely for G-compass such as ultraconserved regions and transcription factor binding profiles. While G-compass provides such biological resources at the genomic level, the alignments of G-compass are also available for download, which will be of great help for the functional annotations and evolutionary analysis based on comparative genomics. As an example, in order to reveal the evolutionary events that occurred in genomic sequences, we examined the alignments of human and apes. We found that the adjacent regions to gaps have higher nucleotide substitution rates than the other regions, and that these regions show different nucleotide substitution patterns resulting in higher AT, AC, TG mismatches. One of the possible explanations of this trend would be small-scale inversion in the alignments. In the human-chimp genome alignments, we found 19,800 candidates of small-scale (<146 bp) inversions, suggesting that not a few small-scale inversions had occurred during human evolution. This observation indicates that the comparative analysis using the alignments of G-compass would improve our understanding of the genome evolution.

Evola: A Database of Orthologous Genes and Gene Families between Human and Other Vertebrates Detected by Whole Transcriptome Analyses

Ryuichi Sakate1,2, Yuichiro Hara1, Akihiro Matsuya1,3, Jun-ichi Takeda1, Takuya Habara1, Chisato Yamasaki1, Takashi Gojobori1,4, Tadashi Imanishi1

1Biological Information Research Center, AIST, Tokyo, Japan, 2Department of Disease Bioresources Research, NIBIO, Osaka, Japan, 3Hitachi Ltd., Tokyo, Japan, 4Center for Information Biology and DNA Data Bank of Japan, National Institute of Genetics, Shizuoka, Japan

Evola is a database of originally detected orthologs between human and other 13 vertebrates (from primates to fishes). Orthologs are genes in different species that evolved from a common ancestral gene by speciation. Ortholog information is indispensable in various fields of life science such as drug discoveries and iPS/ES cell researches where mice and other species are used as human model. The gene sets analyzed in Evola were compatible with H-InvDB, an integrated database of annotated human genes (http://hinv.jp/), that aims to provide the most comprehensive set of human genes by mapping and clustering whole transcripts of DDBJ, Ensembl and RefSeq. Orthologs between human and other vertebrates were detected by constructing pairwise genome alignments, by comparing transcripts mapped onto them by exon by exon (to distinguish alternative splicing variants), and by phylogenetic tree-based curation. Evola also provides duplicated gene family (DGF) information between human-chimp, macaque, mouse and rat. A DGF was defined to be an one-to-one orthologous gene cluster consisting of >1 genes between human and other vertebrates. Single-linkage clustering of human representative transcripts (one transcript per gene) was conducted and the resulting clusters were merged if they contained a common set of many-to-one (human-to-other vertebrates) orthologs. Evola provides ortholog information for 22,496 human genes and 3,257 human DGFs. Evola has been publicly available since 2005. We intended Evola to be an ortholog database not only of known but also of unknown transcripts that are rapidly increasing in number in the era of next generation sequencers. Evola can be accessed for free of charge at http://hinv.jp/evola/.
Development of 12 Genic SSR Markers for a Biofuel Grass, *Miscanthus sinensis* (Poaceae)

Ho, Chuan-Wen, Wu, Tai-Han, Chiang, Tzen-Yuh, Huang, Chi-Chun

Department of Life Sciences, National Cheng-Kung University, Tainan, Taiwan 701

Next generation sequencing platform has been considered as a powerful tool for transcriptome sequencing project of *de novo* organisms. It can also rapidly develop new genic SSR markers for a non-model species that is deficient of whole-genome sequences. *Miscanthus*, a biofuel plant, has been widely applied in Europe and the United States. To select strains with high photosynthesis efficiency and/or stress resistances from wild populations, we conducted large-scale transcriptome sequencing of *Miscanthus sinensis* var. *glaber* at a vegetative stage with a 454 sequencer. A total of 20,174 contigs from 241,051 reads of 454 sequencing were screened for di- to hexanucleotide. In total, 298 contigs containing SSR markers were identified. Twelve polymorphic SSR markers in *M. sinensis* were characterized with a number of alleles ranging from 6 to 13 per locus, and total observed and expected heterozygosities ranging from 0.33 to 0.94 and from 0.53 to 0.83, respectively. Cross-species transferability was identified as all loci can be applied to its closely related species, *Miscanthus floridulus*. The application of genic SSR markers may provide a tool for assessing the genetic diversity of functional traits and detecting the potential genes of local adaptation. In total, three loci, i.e., locus 2663, 5417 and 15763, were identified under positive selection.

Profile on Human Druggable Space Revealed by Integrated Analyses of Bioinformatic and Chemoinformatic Resources

Yasuyuki Fujii¹, Takashi Gojobori²³, and Tadashi Imanishi²

¹Innovation Center Okayama for Nanobio-targeted Therapy, Okayama University, Japan, ²Biomedical Information Research Center, National Institute of Advanced Industrial Science and Technology, Japan, ³Center for Information Biology and DDBJ, National Institute for Genetics, Japan

Most of pharmaceutical effects exhibited by modern drug compounds are explained based on interaction with gene products related to disordered biochemical networks. The genes that encode proteins to which approved drugs or drug-like compounds are referred to as “druggable genes”. Profiling the annotation of druggable genes should be effective for narrowing down the druggable space, and may contribute to improve our strategy for drug discovery. In this study, we aimed at profiling biological features of druggable genes by use of curation for human genes and a chemoinformatics resource. The drug-gene association (DGA) archived in DrugBank 2.1(http://www.drugbank.ca/) was utilized for defining current druggable space. The annotation of human genes was obtained from H-InvDB release 6.2 (http://hinv.jp/), because it covers a wide range of human OMICs data. These data were merged by similarity search between protein sequences in H-InvDB and DrugBank. For profiling the druggable genes, a web-tool for gene-set enrichment analysis called HEAT (http://hinv.jp/HEAT/) was utilized. As a result, 1,086 genes were found to be targets of at least one of FDA-approved drugs, and HEAT analysis of these genes successfully profiled the current human druggable space. Based on GO terms, gene products of 213 druggable genes were localized in membrane (GO:0016020, P=1.22E-83) and the 57 genes encoded GPCR superfamily. 108 genes exhibited catalytic activity (GO:0003824, 1.57E-70), which means that nearly 26% of all human enzymes are successful drug targets. A data-mining using H-InvDB revealed that 87% of druggable genes have one or more nonsynonymous SNPs (nsSNPs). For example, 197 nsSNPs were found in phenylalanine-4-hydroxylase (HIT000220124), a target for anti-Parkinson drugs. The dosage control of these agents should be more carefully evaluated considering the nsSNPs.
**Arabidopsis thaliana: further Exploiting This Plant Reference Genome**

Mara Sangiovanni¹, Alessandra Vigilante¹ and Maria Luisa Chiusano¹

¹Department of Soil, Plant, Environmental and Animal Production Sciences, University of Naples Federico II, Via Università 100, 80055 Portici, Italy

_Arabidopsis thaliana_ is the reference in plant genomics since its complete genome sequence was the first one to be made available in 2000. As a reference model the _Arabidopsis_ genome should be fully reliable and safely annotated; moreover its organization should be well understood in terms of evolutionary mechanisms that gave rise to the actual genome structure. However the presence of widespread intragenome duplications, together with the loss of many gene copies, associated to possible ancient and recent polyploidization events, really complicates the interpretation of the factors contributing to the genome shaping, thus limiting also the role of this genome as a reference in plant comparative genomics.

To further exploit the information available, and with the aim of supporting the genome annotation of other plant and crop species, we investigated the organization of the _Arabidopsis_ genome in terms of paralogous genes. We identified all the possible pair-wise similarities between genes classifying structurally related ones into networks, with each gene belonging to only one network given the presence of one or more paralogy relationships. The organization of the genome into networks of duplicated genes therefore provides a novel view for intra-genome and inter-genome comparative analyses, also permitting an appropriate investigation of gene families evolution. We also focused on the identification of single copy genes (singleton), because their presence in a highly duplicated genome is still an intriguing evolutionary issue. Furthermore the analysis of singleton genes revealed some flaws into the annotation data of _Arabidopsis_ genome further indicating the need of a more accurate annotation process.

This work is supported by the Agronanotech Project (Ministry of Agriculture, Italy)

---

**MAIZEGDB.ORG, the Maize Genetics Cooperation and the 2500 MB B73 Genome-Generated Tsunami**

Mary Schaeffer*,¹,², Jack Gardiner*,¹,², Lisa Harper*,¹, Carson Andorf², Darwin Campbell¹, Ethy Cannon⁵, Taner Sen¹,², Carolyn Lawrence §¹,²

¹USDA-ARS USA, ²University of Missouri USA, ³University of Arizona USA, ⁴UC Berkeley USA, ⁵Iowa State University USA, *Curator, §Director

Advances in sequencing technology have made it possible to sequence the 2500 MB B73 maize genome, both cheaply and in a relatively short time. Nearly simultaneously, other sequencing-based data are on the leading edge of a data tsunami: sequenced differences (currently >300,000 SNP for >1000 inbred lines and related species) and gene expression data. The MaizeGDB team integrates access to these datasets after considerable cooperation among both data users and suppliers, and typically while the data are being generated. In this way, data suppliers have resources in place to format data ready for integration with MaizeGDB. And, we learn what details, including links to other resources, are needed to support our user community and what interface development is required. An interesting result is that we inadvertently serve as outreach for the data suppliers, explaining the caveats and expected availabilities of data not yet at MaizeGDB. Contact with users is at conferences, by online tutorials, personal visits, or by surveys from the Maize Executive Genetics Committee addressed to all persons listed in MaizeGDB as a Cooperator.
In addition to maintaining a rice gene list, Oryzabase, a database of rice science, has also been hosting a website to submit newly identified genes. The current rice gene dictionary is based on a traditional rice trait gene dictionary and extended through time with new genes which are manually retrieved from journal articles. Currently over 4000 genes are listed. However it is still insufficient in both quantity and quality.

Recently we started to apply automated extraction before manual annotation to enhance the efficiency of gene extraction. To improve the quality of information, we started collecting relevant DNA accessions and LOCUS-ID. In Oryzabase, we try to employ the help of users to overcome this information insufficiency by adding a comment box in each gene page to allow researchers to write additional information. Besides adding GO (gene ontology) and TO (trait ontology) to the genes, PO (plant ontology) has been newly assigned to enable better semantic query processing. The current status of gene extraction and ontology assignment and its related problems will be introduced.

ChEBI, an Open-access Chemistry Resource for the Life Sciences: Facilities for On-line Submission and Curation

Marcus Ennis, Nico Adams, Adriano Dekker, Paula de Matos, Janna Hastings, Kenneth Haug, Duncan Hull, Zara Josephs, Gareth Owen, Steve Turner and Christoph Steinbeck

European Bioinformatics Institute, Wellcome Trust Genome Campus, Cambridge, UK

ChEBI (Chemical Entities of Biological Interest) is a database of ‘small’ molecular entities structured around a chemical ontology. It contains almost 600,000 entries, of which approximately 20,000 have been manually curated, as well as entries for groups (parts of molecular entities) and classes of entities. It provides a wide range of information such as chemical nomenclature, structures and related chemical values, and establishes interrelationships between entities in the ontology, in terms of both structure and role. ChEBI places a strong focus on quality, with exceptional efforts being applied to upholding IUPAC nomenclature recommendations and best IUPAC practices when drawing chemical structures.

To invite the community to participate more directly in the future growth and development of ChEBI, we have developed a web-based software utility to enable direct user submissions. Users are encouraged to carry out as much of their own manual curation as possible, e.g. by adding multiple synonyms and database cross-references, and by creating multiple relationships within the ontology. The submissions are automatically validated for uniqueness (both of name and chemical structure) and correctness (such as checking that no non-allowed cycles have inadvertently been created in the ontology graph structure, and that the ontology relationships which have been specified are allowed between entities of the relevant types). Once a submission has passed the required validations, it is submitted to the ChEBI database, at which time it receives its unique ChEBI identifier. It will then become visible to the public (as a preliminary entry) as part of the monthly ChEBI release. To date, ChEBI has received over 750 such external submissions.

Knowledge Discovery and Database Construction for Genes Contributing to "Sustainable World" and "Human Health" & TRNA Gene DataBase Curated Manually by Experts under Collaboration with Undergraduate Students and Senior Scientists.

Hiroshi Uehara1, Toshimichi Ikemura1, Hachiroy Inokuchi1, Chieko Wada1, Yuko Yamada1, Akira Muto2, Yasuo Ohara1, Takashi Abe1

1Nagahama Inst. of Bio-Sci. and Tech., 2Hirosaki Univ., JAPAN

Database for Genes Contributing to "Sustainable World" and "Human Health" has been constructed as an educational program for undergraduate students. In the program, a student first chooses target genes at one’s own initiative, and the student retrieves genes of attention from a vast amount of genomic sequences, which have been obtained by metagenomic analyses of environmental samples and stored in the International Nucleotide Sequence Databases Consortium with no functional and phylogenetic annotation. When candidates for the useful genes were found, the genes were registered in our Database for publication with functional and phylogenetic annotations along with the student’s name (http://dbcls.nagahama-i-bio.ac.jp/).

Senior expert scientists have curated the newly-found data under collaboration with students.

The tRNA Gene DataBase Curated manually by Experts "tRNADB-CE"1 was constructed, under collaboration between students and young and senior scientists. The 287,103 tRNA genes in total were registered in this database, which was manually curated by senior scientists to obtain a high-quality database.

These databases have been constructed as a training program of the "Integrated Database Project in Life Science Fields" supported by the Ministry of Education, Culture, Sports, Science and Technology, Japan.

Incorporating Community Annotation Interfaces into the CIPRO2.5 Database with Comprehensible Sketches to Support Quick Annotations of Proteome Data.

Keisuke Ueno1, Kouki Yonezawa1 Katsuhiko Mineta1, Toshinori Endo1

1Graduate School of Information Science and Technology, Hokkaido University, North 14 West 9, Sapporo, Hokkaido 060-0814, Japan

User annotation or comment interfaces are now widely used in several web sites such as journals, news, weblogs and Wikipedia. However, there are only a few biological databases with annotation interfaces. The *Ciona intestinalis* protein database (CIPRO) was created in order to provide integrated proteome data especially for experimental biologists. The current database contains 89,673 unique sequences covering all the known and predicted gene models. Typical tasks include which gene models are reliable and which function is plausible. The human-curated annotation is most important for the meaningful database.

Here we incorporate three new functions into the CIPRO2.5 database (http://cipro.ibio.jp/2.5/), providing enriched resources for the users. First, a community annotation interface as web forms and a user comment editor with rating its comment were added. Second, the web pages were specifically designed to compact for quickly understood overviews. For example, cytolocalization was automatically provided by a color-depicted cell image based on the intensity instead of the numeric values of raw data. The expression data of EST, microarray and 2D-PAGE were integrated as one chart. In addition to these data, the images of transmembrane prediction, domain and motif search, and the OMIM ortholog on the chromosome map were included in each protein page. The last, even the BLAST and PMF search were added to combined fields in the retrieval system.

As a result, a total of 11,134 pages were annotated by our community. Furthermore, 2,186 comments were added to the database. Those annotated data are freely accessible at the CIPRO2.5 web site.
H-InvDB: A Comprehensive Annotation Resource For Human Transcriptome

Chisato Yamasaki¹, Jun-ichi Takeda¹, Takuya Habara¹, Makoto Ogawa²,³, Akiko Noda¹, Ryuichi Sakate¹, Katsuhiko Murakami², Tadashi Imanishi¹ and Takashi Gojobori¹,⁴

¹Biomedicinal Information Research Center, National Institute of Advanced Industrial Science and Technology, Japan
²Japan Biological Informatics Consortium, Japan
³DYNACOM Co., Ltd., Japan
⁴National Institute of Genetics, Japan

H-Invitational Database (H-InvDB: http://www.h-invitational.jp/) is a comprehensive annotation resource for human transcriptome. By extensive analyses of all human transcripts, we provide curated annotations of human genes, transcripts and proteins that include gene structures, alternative splicing isoforms, non-coding functional RNAs, protein functions, functional domains, sub-cellular localizations, metabolic pathways, protein 3D structure, genetic polymorphisms, relation with diseases, gene expression profiling, molecular evolutionary features, protein-protein interactions (PPIs) and gene families/groups. The latest release of H-InvDB (release 7.0) provides annotation for 296,912 human transcripts in 46,499 human gene clusters based on human full-length cDNAs, mRNAs and the reference human genome sequences (NCBI b37.1). H-InvDB consists of three main views, the Transcript view, the Locus view and the Protein view, and six sub-databases; G-integra, H-ANGEL, DiseaseInfo Viewer, Evola, PPI view and Gene Family/Group view. We also provide data mining tools such as “Navigation search”, an extended search system that enables complicated searches by combining 16 different search options (http://www.h-invitational.jp/hinv/c-search/hinvNaviTop.jsp) and “H-InvDB Enrichment Analysis Tool (HEAT)”, a data mining tool for automatically identifying features specific to a given human gene set (http://hinv.jp/HEAT/).

Data Integration Model and GUIs Used in Human Genome Network Platform

Kazuho Ikeo¹, Takashi Gojobori¹, Genome Network Platform Team

¹National Institute of Genetics, Center for Information Biology and DNA Data Bank of Japan, Japan

The authors introduce the database and Graphical User Interfaces (GUIs) of Genome Network Platform and its application to comprehensive analysis on transcription network. Genome Network Platform has been developed by National Institute of Genetics as a research infrastructure for Genome Network Project (GNP), which is headed by Ministry of Education, Culture, Science and Technology. This platform is designed to provide integrated data that helps elucidating biological networks including transcriptional regulation. There are accumulating data produced by GNP consortium members; genomic information such as transcription start sites and transcription regulatory region, transcriptomic information such as expression and splice variants, proteomic information such as domain structures and protein-protein interactions. Genome Network Platform contains these data produced by GNP consortium members and public data. It will be useful to integrate these data successfully for the researchers in areas including a comprehensive analysis of biological process, but we have some difficulties to integrate and utilize simply the different type of data such as genome, transcriptome and proteome. Moreover the amount of data is too huge to handle with ease, something of summarize is required to handle in all. The authors propose the data integration model and GUIs that are focus on the gene transcription. Here, we explain the model and demonstrate the values of GNP GUIs for study on transcription network.
POPCorn: A PrOject Portal for corn

Ethalinda KS Cannon, Scott M. Birkett, Bremen L. Braun, Taner Z. Sen, Carolyn J. Lawrence

Dept. of Genetics Development & Cell Biology, Iowa State University, Ames, Iowa 50011-3260, USA
Corn Insects and Crop Genetics Unit, USDA-ARS, Iowa State University, Ames, Iowa 50011-3260, USA

The impressive growth of maize data and information available on the Internet means that locating and searching for relevant data and information is increasingly difficult for maize researchers. Finding appropriate sites then learning how to navigate the sites can be difficult and time-consuming. Finding information associated with a particular sequence is likewise challenging, sometimes involving BLASTs at multiple sites and database searches that require in-depth knowledge of the database architecture and other technical information. In addition, data generated by projects can be lost after the completion of the project, or left untouched, only to degrade over time. The POPcorn project is addressing these problems by providing a set of search utilities for locating maize projects and online resources, BLASTing against multiple sequence targets from multiple sources, and for searching sequence-indexed data: data linked to nucleotide and/or protein sequences. In its next stages of development, POPcorn will also provide pipelines and processes for migrating project data to MaizeGDB for long-term storage.

Funding acknowledgement: National Science Foundation (NSF), United States Department of Agriculture (USDA).

Towards Developing Common Standards for Genome Sequence and Annotation.

Tatiana Tatusova, Ilene K Mizrachi

National Center for Biotechnology Information, National Library of Medicine, National Institutes of Health, Bethesda, MD USA

Genome sequencing has proliferated to the extent that it has become pedestrian. However, providing quality annotation for genome sequences has not progressed with sequencing technology. As the ability to perform DNA sequencing of a genome becomes more readily available to the biological community, the tools to provide quality structural and functional annotation for these genomes have not kept up. Many annotation pipelines use a combination of automatic genome prediction methods and manual curation procedures which leads to the variable quality genome annotation in the public databases, partly due to a paucity of standards and reporting. There is a serious need to develop common standards for genome submission and annotation among major sequence databases, large sequencing centers, model organism and curated databases. This past spring, National Center for Biotechnology Information (NCBI) organized an international Genome Annotation Workshop (http://www.ncbi.nlm.nih.gov/genomes/AnnotationWorkshop.html) to address problems with genome annotation and move towards developing common set of standards. The meeting culminated in agreement over a set of guidelines and minimal standards for microbial genome annotation. Of particular note, was the agreement to conform to common protein naming conventions including a subset applicable specifically to microbial proteins. The adoption of common standards by the databases and scientific community will simplify the exchange of knowledge between different resources, and improve manual searching, automatic data-mining, and other functions that rely on consistent naming and/or unique identifiers. NCBI now hosts a set of annotation tools and validation checks for genome submitters to use to check their annotation prior to submission to the International Nucleotide Sequence Database (INSDC). We are committed to improving tools for submitters and developing reporting metrics as a means to improving the genome annotation in the public databases.
Literature Curation of Protein Interactions: Measuring Agreement Across Major Public Databases

Andrei L. Turinsky¹, Sabry Razick²,³, Brian Turner¹, Ian Donaldson²,⁴ and Shoshana J. Wodak¹,⁵

¹ Molecular Structure and Function program, Hospital for Sick Children, Toronto, Canada
² The Biotechnology Centre of Oslo, University of Oslo, Norway
³ Biomedical Research Group, Department of Informatics, University of Oslo, Norway
⁴ Department of Molecular Biosciences, University of Oslo, Norway
⁵ Department of Molecular Genetics, Department of Biochemistry, University of Toronto, Canada

Literature curation of protein interaction data faces a number of challenges. Although curators increasingly adhere to standard data representations, the data that various databases actually record from the same published information may differ significantly. Some of the reasons underlying these differences are well known, but their global impact on the interactions collectively curated by major public databases, has not been evaluated.

Here we quantify the agreement between curated interactions from 15,471 publications shared across nine major public databases. Results show that on average, two databases fully agree on 42% of the interactions and 62% of the proteins curated from the same publication. Furthermore, a sizable fraction of the measured differences can be attributed to divergent assignments of organism or splice isoforms, and to alternative representations of multi-protein complexes. Our findings highlight the impact of divergent curation policies across databases, and should be relevant to both curators and data consumers interested in analyzing protein-interaction data generated by the scientific community. Availability: http://wodaklab.org/iRefWeb

Neuroscience Resources and the Neuroscience Information Framework

Anita Bandrowski, Jeff Gerthe, Timothy Tsui, Maryann Martone on behalf of the NIF project team

¹University of California, San Diego, USA

High scientific quality data and materials are not always ranked very high on commercial search engines. Additionally, many such materials are maintained within databases that are inaccessible to search technology. To solve these problems, the Neuroscience Information Framework (NIF: http://www.neuinfo.org) was created to aid the neuroscience community to discover useful digital resources, such as academic databases, and it has also developed a large digital catalog of resources that are related to neuroscience.

NIF has developed a “resource ontology”, and synchronized to a large extent the effort with the BRO (biomedical resource ontology) and assigned all 3K+ digital resources within the catalog to one or more of these ontological categories, making it possible to discover for example: all atlases through the NIF catalog. The resource categories are: Data, Funding, Job, Material, People, Services, Software, and Training. Most of these contain sub-categorizations, definitions and synonyms and they can be viewed in the Neurolex found at http://neurolex.org/wiki/Resource_Type_Hierarchy. Each “resource descriptor” has been coded in OWL format files and has a unique identifier as well as synonyms and subclasses to help search systems locate data.

Cataloging efforts of digital resources are tricky because unlike publications they can change at any time, including major shifts of data and structure, therefore the curation effort must be reasonably scoped and relatively nimble. NIF has developed some self reporting tools in the DISCO suite, accessible from http://disco.neuinfo.org/webportal/discoDashboardShow.do, to create just such a nimble resource identification system, and allows for import and export with the BiositeMaps schema, so that a resource provider does not have to fill out metadata in multiple formats.
Glycome informatics has started to develop during this past decade, especially with the development of glycan structure databases by KEGG and others in the United States and Europe. However, with the development of various databases, many different formats representing glycan structures have emerged. Recently, the Consortium for Functional Glycomics (CFG) and EuroCarb have collaborated to develop a glycan structure standard for data exchange, based on the GLYDE-II xml representation for carbohydrates and glycoconjugates. In addition, web services have been developed in many glyco-databases to enable the exchange of such data. As a first attempt for glycan data standardization, an xml-based standard format was developed, along with several interfaces such that basic query searches can be implemented with a common standard interface. These interfaces include: retrieval of glycan information based on database ID, glycan substructure, mass and glycan composition. The retrieved data can also be used for operational queries such that more complex queries can be made. The implemented services in EuroCarb and CFG will in turn be registered in BioMoby such that workflows can be generated based on them. This will also potentially allow the incorporation of glyco-data with other related databases. Although this area of glyco-informatics is still in the early stages, many useful data resources and web services have been developed, and these will be presented.

SciNeS Search: Inference Search over an Integrated Life-sciences Database Based on the Semantic Web

Norio Kobayashi¹, Manabu Ishii¹, Yuko Yoshida¹, Yuko Makita¹, Akihiro Matsushima¹, Yoshiki Mochizuki¹, Tetsuro Toyoda¹

¹Bioinformatics And Systems Engineering (BASE) division, RIKEN, Japan

Integration of various omics databases is a fundamental approach for comprehensive understanding of life phenomena. SciNeS is a Semantic Web-based data repository that realises database integration by collecting metadata of databases described as a set of triples consisting of two bio-items such as gene and ontology terms and a semantically defined relationship between these bio-items. These triples are curated from various life-sciences databases including gene-publication relationships manually curated from MEDLINE and gene-phenotype relationships extracted from published databases by computer programs. SciNeS internally forms a metadata network by maximally concatenating triples. SciNeS Search realizes a cross-search over the integrated databases by traversing the metadata network managed by SciNeS. SciNeS Search allows a user to discover bio-items included in the ontological category specified by the user. The results include bio-items not only having a user’s keyword but also semantically inferred from these bio-items using metadata relationships, and are ranked using statistical scores computed on the basis of the metadata network. SciNeS Search is available via a user’s web browser and provides search service against over 30 million bio-items of 100 thousand categories and 100 million relationships curated from over 100 public omics databases of various species including human, mouse, rice and Arabidopsis.

SciNeS http://database.riken.jp

Metabolic Pathway Integration of *Vibrio vulnificus* CMCP6

Kwangjoon Jeong\(^1,2\), Joo Young Lee\(^1\), Joon Haeng Rhee\(^1,2\)

\(^1\)Department of Microbiology, Chonnam National University Medical School, Korea, Republic of
\(^2\)Clinical Vaccine R&D center, Korea, Republic of

*Vibrio vulnificus* is the gram negative estuarine bacterium which cause severe septicemia and death for whom was susceptible to it. Based upon the genome sequences of *V. vulnificus* CMCP6, we constructed the optimized in-house genome information DB (CVAS, Clinical Vaccine R&D Center Analysis System) targeting efficient analysis for virulence studies. The importance of bacterial metabolism in host environment has been esculted by recent wet bench studie. Those *in vivo* expressed bacterial pathways are highlighted as new targets for better understanding of pathogenesis and developing new therapeutic modalities. Here we report our design and strategy for the integration system of our genome DB with the pathway data from the KEGG (Kyoto Encyclopedia of Genes and Genomes). Using this system, we could analyze more intensively *in vivo* gene expression pattern of *V. vulnificus*. We hope this could set a turning point of the computerized modeling of the host-*V. vulnificus* interaction and provide an efficient tool for searching the new therapeutic modalities against *V. vulnificus* infection.

Challenges And Solutions To Rice Biological Data Curation at IRRI

Richard Bruskiewich\(^1\), Ramil Mauleon\(^1\), Victor Jun Ulat\(^1\), Rosemary Shrestha\(^2\), Kenneth McNally\(^1\), and Chengzhi Liang\(^1\)

\(^1\)International Rice Research Institute (IRRI), Philippines
\(^2\)Centro Internacional de Mejoramiento de Maíz Y Trigo (CIMMYT), Mexico

The International Rice Research Institute (IRRI) hosts an ambitious scope of rice germplasm research generating a diversity of rice biological data including structural and functional genomics, breeding and phenotypic data. Systematic documentation and publication of such data, cross-linked to related information in public online international plant databases is a major institutional challenge.

In this poster, we give a broad overview of data sets under curation at IRRI, highlighting key curatorial issues including semantic codification (“rice ontology”), database representation, analysis protocols and end user tools. In particular, we will show the ongoing evolution of germplasm genotype and phenotype data, in its relation to genetic and genomic information available from public plant genome projects, by direct comparison to the rice genome, or indirect comparative analysis to other plant species.
Integral Presentation Of Experimental Results And Bioinformatic Analyses On Ciona intestinalis Proteins

Kouki Yonezawa¹, Keisuke Ueno¹, Katsuhiko Mineta¹, and Toshinori Endo¹

¹Graduate School of Information Science and Technology, Hokkaido University, Japan.

For the aim of facilitating biologists to find new knowledge through the comprehensive proteomic and transcriptomic data of Ciona intestinalis, we have been developing an intuitive and integral view of experimental and bioinformatic results for each protein, which, unlike on human proteins, was incomplete for Ciona proteins. In the presentation, we integrated in a view the experimental results of 2D-PAGE analyses done at Shimoda Marine Research Center, University of Tsukuba, 3D-view of expressions (3DPL) done by Hotta et al., and cross-references with JGI version 1 and ENSEMBL. A total of 1,673 entries of 2D-PAGE and 57 of 3DPL are currently contained in CIPRO database. The information is summarized in a single page for a protein. Together with BLAST results, comparative genomics, expression profiles, domain search results, localization, phylogeny etc., we provide biologists the highly integrated view amongst the existing databases, which facilitates to understand the protein. For example, the 2D-PAGE and the microarray results of the same protein are sometimes different, which implies the spatial or temporal gaps of the expressions between the mRNA and the protein. Besides, two functionalities are implemented. One is an improved 2D-PAGE viewer, which provides the compact view of spots and experimental information, and another is peptide fragment mass search functionality as improved PerMS algorithm by Hozumi et al., by enabling detection of seven kinds of modifications and suspected contaminants.

The Neuroscience Information Framework: A Unified Semantic Framework for Discovery and Integration of Biomedical Data and Resources on the Web

J. S. Grethe¹, A. Bandrowski¹, F. Imam¹, A. Gupta¹, A. Memon¹, G. M. Shepherd², P. Miller², L. Marenco², D. Van Essen³, E. Reid³, G. Ascoli³, S. Polavaram³, P. R. Sternberg⁵, H.-M. Mueller⁵, A. Rangarajan⁵, M. E. Martone¹

¹Ctr. for Res. in Biol. Systems, UCSD, USA;²Neurobio., Yale Univ., USA;³Anat. and Neurobiolog, Washington Univ., St Louis, USA;⁴Krasnow Inst. for Advanced Study, George Mason Univ., USA;⁵Caltech, USA

Informatics and new web technologies are becoming increasingly important to biomedical researchers. An initiative of the NIH Blueprint for Neuroscience Research, the Neuroscience Information Framework (NIF; http://neuinfo.org) enables discovery and access to public research data, contained in databases and structured web resources (e.g. queryable web services) that are sometimes referred to as the deep or hidden web, and resources through an open source dynamic inventory of biomedical resources that are annotated and integrated with a unified system of biomedical terminology. The NIF Database Federation (with more than 60 independent databases) allows for direct search, discovery and integration of database content. Search and annotation of resources and resource content is enhanced through the utilization of a comprehensive ontology (NIFSTD; http://purl.org/nif/ontology/nif.owl), built as a set of modular ontologies. To enable broad community contribution to NIFSTD, NeuroLex (http://neurolex.org) is available as a wiki that provides an easy entry point for the community. New services being provided include a complete full text index of PubMed Central’s Open Access articles, an annotation framework (allowing content from the NIF data federation to be annotated by the NIFSTD ontologies) that allows data resources to be efficiently indexed (Lucene) for user searches and also made available as a SPARQL (RDF) end-point, entity highlighting services, and informational pop-ups (NIF Cards) that can be linked into any application. NIF cards draw upon NIFSTD and the NIF data federation to display information about an entity and provide customized search options depending upon the domain. As the NIF Cards evolve, they will provide the basis for linking results into the large ecosystem of linked data. In addition, enhancements have been made to the NIF ontologies, including the ability to intelligently handle inferred classes such as “GABAergic neuron”.

89
Wormbase Curation Interfaces And Tools

Xiaodong Wang, Paul Sternberg and the WormBase Consortium,
Division of Biology, California Institute of Technology, Pasadena, CA, USA

Curating biological information from the published literature can be time- and labor-intensive especially without automated tools. WormBase\(^1\) has adopted several curation interfaces and tools, most of which were built in-house, to help curators recognize and extract data more efficiently from the literature. These tools range from simple computer interfaces for data entry to employing scripts that take advantage of complex text extraction algorithms, which automatically identify specific objects in a paper and presents them to the curator for curation. By using these in-house tools, we are also able to tailor the tool to the individual needs and preferences of the curator. For example, Gene Ontology Cellular Component and gene-gene interaction curators employ the text mining software Textpresso\(^2\) to indentify, retrieve, and extract relevant sentences from the full text of an article. The curators then use a web-based curation form to enter the data into our local database. For transgene and antibody curation, curators use the publicly available Phenote ontology annotation curation interface (developed by the Berkeley Bioinformatics Open-Source Projects (BBOP)), which we have adapted with datatype specific configurations. This tool has been used as a basis for developing our own Ontology Annotator tool, which is being used by our phenotype and gene ontology curators. For RNAi curation, we created web-based submission forms that allow the curator to efficiently capture all relevant information. In all cases, the data undergoes a final scripted data dump step to make sure all the information conforms into a readable file by our object oriented database

1 http://www.wormbase.org
2 http://www.textpresso.org

RDF Curator: A Novel Workflow that Generates Semantic Graph from Literature for Curation Using Text Mining

Yusuke Komiyama\(^1\), Osamu Gotoh\(^1\)

\(^1\)Department of Intelligence Science and Technology, Graduate School of Informatics, Kyoto University, Japan

Background: There exist few databases that enable cross-reference among various research fields related to bioenergy. Cross-reference is highly desired among bioinformatics databases related to environment, energy, and agriculture for better mutual cooperation. By uniting Semantic Graph, we can economically construct a distributed database, regardless of the size of research laboratories and research endeavors.

Purpose: Our purpose is to design and develop a workflow based on RDF (Resource Description Framework) that generates Semantic Graph for a set of technical terms extracted from documents of various formats, such as PDF, HTML, and plain text. Our attempt is to generate Semantics Graph as a result of text mining including morphological analysis and syntax analysis.

Result: We have developed a prototype of workflow program named “RDF Curator”. By using this system, various types of documents can be automatically converted into RDF. “RDF Curator” is composed of general tools and libraries so that no special environment is needed. Hence, “RDF Curator” can be used on many platforms, such as MacOSX, Linux, and Windows (Cygwin). We expect that our system can assist human curators in constructing Semantic Graph.

Conclusion: Although fast and high throughput, the accuracy of the present version of “RDF Curator” is lower than that of human curators. As a future task, we have to improve the accuracy of the workflow. In addition, we also plan to apply our system to analysis of network similarity.
InterPro Curation: Integrating Predictive Protein Signatures Into Biological Hierarchies

Sarah W Burge, Louise Daugherty, David Lonsdale, Craig McAnulla, Siew-Yit Yong, Prudence Mutowó, Amaia Sangrador and Sarah Hunter.

1EMBL-EBI, The Wellcome Trust Genome Campus, Hinxton, CB10 1SD, UK

InterPro is an integrated database of predictive protein signatures used for the classification and automatic annotation of proteins and genomes. As InterPro curators, we are responsible for assimilating information from our member databases and communicating it to our end users in a way that adds value to each individual signature. We categorise signatures according to their type (for example, Family, Domain or Repeat) and annotate entries with links to other databases, abstracts and protein matches.

The InterPro database also identifies relationships between entries. For example, signatures at a general Family level are related to more specific subfamilies through a Parent/Child relationship. Families may also Contain individual Domains. In this manner, we aim to build up a hierarchy of InterPro entries that correctly represents relationships between biological families and domains. Users may then easily identify related proteins and signatures as the InterPro database attempts to map out biological hierarchies. Here we discuss InterPro relations, the criteria for their formation and how they may be useful to users. We will also discuss the challenges of representing biological hierarchies when automating relationship formation and the role manual curation plays in ensuring that we accurately represent biological networks.

Development of Cooperative Work Tools and Workflow for Multimodal Genome Annotation


1Database Center for Life Science, Japan, 2Kazusa DNA Research Institute, Japan, 3Kyoto Saga University, Japan, 4Chuo University, Japan, 5DNA Data Bank of Japan, Japan.

The focus of biological research is rapidly shifting from individual genes to genome-wide systems. To understand the genome-wide data sets, biologists must be able to connect with research fields outside their core competence. In this context, comprehensive genome annotation is very important as infrastructure of genome-wide research. However, annotation is still one of the most difficult task in genome sequencing projects and therefore methodological and technological solutions are urgently needed to reduce annotation costs. For this purpose, we have developed a web-based genome annotation tool, KazusaAnnotation (http://a.kazusa.or.jp). We designed two pilot studies to assess the effectiveness of our system for reducing the cost of cooperative genome annotation.

The first project is a geographically distributed annotation model, designated Gene Indexing (GI). GI is a named-entity recognition that manually extracts gene/protein symbols from each section of full papers including tables and figures and connects to PubMed IDs. An average of 5 curators cooperated over a 34 months period to manually annotate over 7,000 full papers of 14 plant and plant-related microbes (Cyanobacteria and Rhizobia). As the result, over 175,000 gene/protein mention tags were obtained. The second project is the manual annotation of a soil microbe of Rhizobium by 30 investigators over a period of 2 weeks. As a result >1,700 functional annotations for a total of 6,948 genes were revised. KazusaAnnotation provides an easy way to access, edit and store annotations over a flexible web interface. Using this web annotation tool, distributed communities of curators can collaborate efficiently and improve the speed and quality of manual annotations.
Ontology-based Tools to Enhance the Curation Workflow

Patricia L. Whetzel, Nigam H. Shah, Natasha F. Noy, Clement Jonquet, Cherie Youn, Paul Alexander, Michael Dorf, Mark A. Musen

Stanford University, USA

In order to effectively search, retrieve, and analyze data oftentimes it is curated and tagged with ontology terms. However, the amount of effort to curate the existing set of data resources is beyond the limits of purely manual curation. We present three ontology-based tools developed by the National Center for Biomedical Ontology to enhance the curation workflow: Ontology Widgets, Notes, and the Annotator. The Ontology Widgets provide a mechanism to use ontologies in Web-based forms without the need to locally parse and store the ontology. The widgets provide a variety of functionality including term autocompletion and ontology visualization. The Ontology Widgets are implemented for all BioPortal ontologies, including those from the OBO Foundry and Unified Medical Language System. The Notes feature of BioPortal allows structured term proposals to be submitted in order to request the addition or modification of a term in an ontology. The term proposals can be added directly via the BioPortal Web interface or programmatically via the Notes Web service. Notification of new Notes and replies are both RSS- and Email-enabled. Once the term curation process is complete, the OWL class or OBO stanza can be generated via the Notes Web service. Finally, the Annotator can be used to automatically process textual metadata to identify ontology terms found within the text. The Annotator can be accessed programmatically via the Annotator Web service and can be used with all BioPortal ontologies. In summary, the Ontology Widgets, Notes, and Annotator provide mechanisms to enhance curation by helping collect annotated data upon data submission, by facilitating ontology term curation, and by tagging unstructured textual data with ontology terms.

The Annotation of Olfactory Receptor Gene Family in H-InvDB

Akiko Ogura Noda1, Chisato Yamasaki1, Yoshihito Niimura2, Takashi Gojobori1,3 and Tadashi Imanishi1

1Biomedicinal Information Research Center, National Institute of Advanced Industrial Science and Technology, Japan

2Medical Research Institute, Tokyo Medical and Dental University, Japan

3National Institute of Genetics, Japan

Olfaction, the sense of smell, is important for animals to find food, identify mates and offspring, and avoid danger. The olfactory perception begins when odorous ligands activate olfactory receptors (ORs) expressed in olfactory sensory neurons of the olfactory epithelium located in the posterior upper part of the nasal cavity. ORs are encoded by the largest gene family in animals and OR gene family includes many pseudogenes. There are several databases of OR genes, but most of them are merely predictions from the genome sequence, because gene expression levels of the OR genes are very low and there are few evidence of mRNA and protein. We carried out manual annotation of OR genes. First, we extracted candidates of OR genes by using homology search toward human genome based on a modified method of Niimura and Nei (2003) and by using keyword search against sequence description. Then, to find experimental validation of these estimated OR genes, we examined 800 publications through SwissProt and RefSeq entries whether there is expressional or protein evidence of the OR genes. We offer a curated set of 382 OR genes from the H-InvDB gene family page (http://www.h-invitational.jp/hinv/genefamily/index_en.cgi). The dataset includes the following information: definition, intact gene or pseudogene, protein evidence, expressional evidence, ID of Niimura and Nei, chromosomal location, and HGNC gene symbol.
The discovery of intrinsically disordered proteins (IDPs) has brought a paradigm change to structural biology. IDPs are those that do not assume any stable 3D structure by themselves under physiological conditions. Some proteins are fully composed of intrinsically disordered (ID) regions while others contain long ID regions. IDPs are involved in crucial biological processes such as signal transduction, transcription control. Typically functional IDPs switch to more ordered states or fold into stable secondary or tertiary structures upon binding to targets, a phenomenon known as coupled folding and binding. ID regions can be inferred from their amino acid sequences. We have developed an ID region annotation system called DICHOT. In contrast to previously available ID prediction programs that merely identify likely ID regions, DICHOT classifies the entire protein sequence into two categories, structural domains (SDs) and ID regions. Application of DICHOT to the human proteome revealed that ID residue-wise constitutes 35%, SD with similarity to PDB structures comprises 52%, while SD with no similarity to PDB structures accounts for the remaining 13%. Novel structural domains, termed cryptic domains, comprise the last category and are good targets of structural genomics.
EMAGE: A Database of Spatially Integrated Gene Expression Data.

Lorna Richardson, Shanmugasundaram Venkataraman, Jeff Christiansen, Peter Stevenson, Nick Burton, Yiya Yang, Duncan Davidson, Richard Baldock.

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

The EMAGE database is a freely available on-line resource of gene expression patterns in the developing mouse embryo. The data contained in EMAGE is sourced from the literature, large-scale screens and via direct submissions from developmental biology labs. The gene expression data is spatially represented, and uses a series of 3D digital models of mouse embryos, at defined stages of development, to store and display the gene expression patterns within the context of the embryo. Annotating the data into and onto these models to produce standardized representations of every gene expression pattern, is a labour intensive process, but one that we feel is justified by the resultant ability to query the database by space, and to cluster the data by spatial similarity to define syn-expression groups.

More recently the availability of fast and effective 3D visualization techniques such as Optical Projection Tomography (OPT) to record gene expression patterns, has created a need for a process by which this full 3D data can be mapped into the 3D space of the models. Such 3D-3D warping has proved to be a technical challenge, but having developed a working method we plan to start including this mapped data in EMAGE very soon.
Allergens in ChEBI – Collaboration with the IEDB

Zara Josephs1, Marcus Ennis1, Steve Turner1, Gareth Owen1, Björn Peters2, Randi Vita2 & Christoph Steinbeck1

1 EMBL-EBI, Wellcome Trust Genome Campus, Hinxton, CB10 1SD, UK
2 IEDB Project, La Jolla Institute for Allergy and Immunology, 9420 Athena Circle, La Jolla, CA 92037, USA

ChEBI background: Chemical Entities of Biological Interest (ChEBI) is a curated database of small chemicals important in biosystems. As well as a description of chemicals, it provides a semantically rich knowledge base, and an internal hierarchy that organises chemicals by their molecular structure types and potential rôles. Following a recent expansion from 20,000 to 500,000 entities, ChEBI is developing automatic curation protocols to maintain the high standards characteristic of the smaller dataset. Simultaneously, the number of manually curated entities is rising.

The ChEBI-IEDB collaboration: The Immune Epitope and Analysis Resource (IEDB) is a project supported by contract from the National Institute of Allergy and Infectious Diseases (NIAID) with the goal of making epitope-related data on infectious diseases and immune disorders freely available to researchers worldwide. The IEDB also houses cutting-edge analytical tools. In June 2009, ChEBI began working with the IEDB on a project aimed at incorporating into ChEBI, by manual curation, a pilot subset of immunologically important chemicals identified as immune epitopes.

The significance of the project: While the incorporated IEDB items (1,200 in the first instance) have substantially enriched ChEBI, the latter’s multiplicity of synonyms, structure tree lay-out and expertise in describing non-peptidic epitopes has been equally useful to the IEDB in facilitating the search process.

Presentation aims and future goals: That collaboration among curators working on different databases can be reciprocally beneficial has been amply demonstrated by the ChEBI-IEDB teamwork described. We aim to present several aspects of this collaboration, including curation strategies, ontology determinations and a sample of the diverse material curated. We plan to continue to refine our rôle in assisting the identification, understanding and utilisation of biologically meaningful chemicals by engaging in further collaborations.

The International Molecular Exchange Consortium

Sandra Orchard1, Henning Hermjakob1, The IMEx Consortium

1 EMBL-EBI, Wellcome Trust Genome Campus, Hinxton, Cambridge, UK

The IMEx consortium is an international collaboration between a group of molecular interaction databases who have agreed to share curation effort. As an integral part of this exercise the databases have developed, and now adhere to, a single set of curation rules when capturing data from both directly deposited interaction data or from publications in peer-reviewed journals. All participating databases adhere to a “deep” curation model, in which full details of an interaction, including experimental methodology and of the participating molecules are captured. Archival journal curation is managed such that each paper is only entered once, by a single database, thus making optimal use of limited curation resources. Currently the dataset is limited to protein-protein interactions but there are plans to expand this to additional molecule types. Interaction records may be searched, using an implementation of the PSICQUIC webservice (http://code.google.com/p/psicquic/) either at the IMEx website (www.imexconsortium.org) or at the websites of participating member databases. All data is available for download in the HUPO PSI-MI XML and MITAB formats and is freely accessible under the Creative Commons Attribution License. This initiative provides the user with a single point of access to a large pool of non-redundant, consistently annotated interaction data, available across a wide variety of species.
**MEDALS: METI Database Portal for Life Science and Renewal-Checker to Keep up to Date Information.**

Akihiro Matsuya1,2, Katsuhiko Murakami1, Kentaro Mamiya1,3, Akihiko Kishimoto1,4, Tadashi Imanishi5, Takashi Gojobori6

1Japan Biological Informatics Consortium., Japan, 2Hitachi Government & Public Corporation System Engineering,Ltd., Japan, 3Hitachi Software Engineering Co., Ltd., Japan, 4Hitachi, Ltd., Japan, 5BIRC, AIST, Japan, 6CIB-DDBJ, NIG., Japan

For the current research of life science, there are a large number of databases (DBs) and analysis software. However most of them are not known for appropriate users. Therefore, creating a handbook of such products is an imperative task. Toward the development of “integrated database” in Japan, we have launched a portal site of such databases, MEDALS (METI Database portal for Life Science) as an activity taken by the Ministry of Economy, Trade and Industry (METI). We have surveyed all recent METI-related national projects and institutes’ activities, then found 158 products recorded. Of them, we have identified 107 working products (62 DBs and 45 analysis tools), and explained them in the manuals in MEDALS. We can manage the renewal information of those products with Renewal-Checker, which was developed in-house. The system makes us notice of “when the product up to date”. We also provide two downloadable software. By viewing MEDALS, users can recognize the relationships among projects and products at a glance, and easily identify databases what they need. We are still increasing the number of the products in the manual. Another outstanding feature of MEDALS is the three original tools called “MEDALS tools”, which include: 1) Hyperlink Management System: an web server for setting hyperlinks to major databases related to human genes and proteins worldwide.; 2) ID converter system: a web tool for converting IDs used in major databases; and 3) PubMedScan: a automatic recommender of newly published PubMed articles.. In MEDALS, the users can cross-search with keywords for a number of DB as well as other huge number of DBs worldwide, in collaboration with Database Center for Life Science (DBCLS). Different from the search system of DBCLS, MEDALS cross-search focuses on the research on human genes and diseases.

**Mapping Bio-databases and Software by Text-mining**

Kentaro Mamiya1,2, Katsuhiko Murakami1,2, Tadashi Imanishi2 and Takashi Gojobori2

1JBIC, Japan
2BIRC, AIST, Japan

Databases (DBs) and software in life science are increasing in their numbers, and the DB contents are getting rich and changing. It is difficult for users to find appropriate DBs or software for their specific purposes. To help finding DBs and understanding the overall activity of the expanding database world, the DB map with bird's-eye views is essential. Here we developed two kinds of database maps: a similarity map and a categorical map. The former approach first determines the similarity between the DBs by analyzing two abstract texts of the published papers corresponding to DBs by a text-mining method "latent semantic analysis (LSA)", which is widely used in text classification. To visualize the relationships among DBs, we mapped all DBs onto two-dimensional space using multidimensional scaling (MDS), while the relationships of DBs in terms of the distance were conserved as much as possible. The map tells us what DBs are similar to each other in terms of various characteristic keywords, such as pathway, genomes and species names. As an advantage of this methodology, new words or concepts can be considered without costly manual data preparation. Therefore this is effective for large scale DB mapping. We applied the method to map the DBs introduced in the Nucleic Acids Research DB issue 2010. The latter approach maps the DBs in a table depending on two features that characterize the DBs. A user selects two features from several annotated features (e.g. "biological species" and "data type"). We used the DBs and features in MEDALS site (METI Database portal for Life Science, http://medals.jp/etop/). Since MEDALS has several unique annotations, such as updated date, institutes, and type of license, the maps provide intriguing views to understand the set of DBs. The two mapping approaches can be also applied to analysis software. The maps can be viewed at MEDALS portal sites (http://medals.jp/etop/).
NamesforLife Semantic Resolution Services for the Life Sciences: Moving Towards an Extensible and Interoperable System for Naming

George M Garrity¹,², Charles Parker¹, Dorothea Taylor¹, Kara Mannor¹, Catherine Lyons¹

¹NamesforLife, LLC, USA and UK
²Michigan State University, USA

A major challenge in bioinformatics, life sciences, and medicine is using correct and informative names. While this sounds simple enough, many different naming conventions exist in the life sciences and medicine that may be either complementary or competitive with other naming conventions. For a variety of reasons, proper names are not always used, leading to an accumulated semantic ambiguity that readers of the literature and end users of databases are left to resolve on their own. This ambiguity is a growing problem and the biocuration community is aware of its consequences.

To assist those confronted with ambiguous names (which not only includes researchers but clinicians, manufacturers, patent attorneys, and others who use biological data in their routine work), we developed a generalizable semantic model that represents names, concepts, and exemplars (representations of biological entities) as distinct objects. By identifying each object with a Digital Object Identifier (DOI), it becomes possible to place forward-pointing links in the published literature, in databases, and vector graphics that can be used as part of a mechanism for resolving ambiguities, thereby “future proofing” a nomenclature or terminology. A full implementation of the N4L model for the Bacteria and Archaea was released in April and encompasses 14264 names, 13831 taxonomic concepts and 13892 exemplars. It is backed by 11456 distinct references. The system is professionally curated and represents a Tier III resource in Parkhill’s view of informatic services. A variety of tools and web services have been developed for readers, publishers, and others and we are incorporating other taxonomies into the N4L data model, as well as adding additional phenotypic, genotypic, and genomic information to the existing exemplars to add greater value to end users.
454 RNA-Seq Data Analysis: What is New and What is Unchanged?

D'Agostino Nunzio1, Alessandra Traini1, Mara Sangiovanni1, Luigi Fruscianti and Maria Luisa Chiusano1.

1 Dept. of Soil, Plant, Environmental and Animal Production Sciences, University of Naples Federico II, via Università 100, 80055 Portici, Italy

For many years the standard method for acquiring transcriptome information on a large scale has been to capture and sequence fragments of messenger RNA (Expressed Sequence Tags) or, alternatively, their entire sequence (full-length cDNAs) using the Sanger technology. Nowadays the ‘RNA-Seq’ approach, which is based on next generation sequencing technologies, is capable of generating much more data per experiment. Because sequencing platforms (e.g. 454 by Life Science, Illumina/Solexa, SOLiD by Applied Biosystem) are undoubtedly different with regard to the number, the length and the quality of the reads they generate, one important issue is to select the most appropriate technology in agreement with the specific scientific target. Otherwise, it is very hard to make sense of thousands of reads, forcing challenges both for the complexity in data management and for obtaining a deeper understanding of biology.

We describe procedures, tools and features of a prototype pipeline for processing 454 reads. It has been tuned for improving performance and effectiveness starting from an existing EST pipeline implemented by the group (D'Agostino et al., 2005). The method here proposed has been applied to analyse and organize specific collections accessible at our website (http://cab.unina.it). In particular, we present the efforts we made to i) evaluate which are the advantages, if any, to combine a clustering and an assembly approach; ii) convert sequence data into functional annotation and iii) make information available to the scientific community. Finally, preliminary results of specific analyses of 454 RNA-Seq collections will be presented to highlight new challenges and opportunities in data management.

This work is supported by the Agronanotech Project (Ministry of Agriculture, Italy)

Analysis Oryza nivara genome via Illumina sequence platform

Cheng-Chieh Wu1, Fu-jin Wei1, Lin-tzu Huang1, Yann-Rong Lin2, Rod A. Wing3, Yue-ile Hsing1

1Institute of Plant and Microbial Biology, Academia Sinica, Taipei, Taiwan
2Department of Agronomy, National Taiwan University, Taipei, Taiwan
3Arizona Genomics Institute, Tucson, Arizona, USA

We have been participated in “The International Oryza Map Alignment Project” (I-OMAP) since 2009; the goal of the project is to generate RefSeqs of several AA genome species (i.e. USA – O. glaberrima, France – O. longistaminata, Australia – O. meridionalis, Taiwan – O. nivara, China – O. rufipogon, Brazil – O. glumaepatula). Oryza nivara is the only Asian annual wild rice and was previously found in Taiwan. We have already prepared the minimum tilling path using BAC end sequences and fingerprinting data. The 12 chromosomes have almost 3,000 BAC clones. About thirty BAC DNAs are pooled together and then sequenced by Illumina paired-end sequence platform. Since the size of each BAC pool is similar to that of the E. coli genome, we should be able to assemble the short read efficiently. In other word, we will provide high-quality reference genome sequences with annotation for wild rice, which can then be used as resequencing templates to capture allelic diversity and structural variation for functional and evolutionary studies in the future.

We attended the International Rice Genome Sequencing Project and worked on sequencing of the chromosome 5. Thus, we start the assembly and annotation of chromosome 5 of Oryza nivara. The analysis strategies include: trimming raw sequence data, assembling clean reads, and mapping reads to genome data. We display the preliminary annotation data via genome browser which illustrates the composition of Oryza nivara with MSU cDNA data and Nipponbare pseudomolecules. For the next step, we will focus on the accurate annotation.
Towards the Draft Genome Sequences of Wild Rice Species: the Landmarks for Future *Oryza* Genomics

Hajime Ohyanagi\(^1\), Eli Kaminuma\(^1\), Toshifumi Nagata\(^1\), Takako Mochizuki\(^1\), Hideki Nagasaki\(^1\), Yasukazu Nakamura\(^1\), Tomoyuki Aizu\(^1\), Atsushi Toyoda\(^1\), Asao Fujiyama\(^1\) and Nori Kurata\(^1\)

\(^1\)National Institute of Genetics, Japan  
\(^2\)Mitsubishi Space Software Co., Ltd., Japan

Taking advantage of the complete genome sequence of *Oryza sativa* ssp. *japonica* cultivar Nipponbare and current high-throughput sequencing technologies, our project has launched with the aim of revealing the history of wild rice genomes and establishing their molecular resources for future wild rice genomics. We employed a high-throughput sequencing technology by illumina GAIIx and are dealing with two wild AA genome species and a CC genome species. As a first step, low-quality reads were excluded with a custom quality filter. For CC genome (distantly related with *Oryza sativa*, AA genome), we are working on multiple libraries of various insert sizes, aiming effective de novo assembly with programs (Velvet and SOAPdenovo). On the other hand, we have generated single library for each AA genome, and employed reference mapping strategy with a program (BWA). As for reference mapping method, the IRGSP genome sequence build5 (*O. sativa*, AA genome) without repeat-masking was used as the reference genome. As well, we collaboratively feed back the accumulated technical know-how to our DDBJ Read Annotation Pipeline (http://p.ddbj.nig.ac.jp/) for its further improvement. The latest status of our project will be presented.
RefEx: Reference Expression Dataset for Functional Curation of Transcriptomes

Hiromasa Ono1, Kosaku Okubo1,2, Toshihisa Takagi1,2, and Hidemasa Bono1

1 Database Center for Life Science, Research Organization of Information and Systems, Japan
2 National Institute of Genetics, Research Organization of Information and Systems, Japan

In Database Center for Life Science (DBCLS) in Japan, we have been tackling the problem how to organize various types of gene expression data and huge amount of transcript sequences. One of the purposes of the gene expression analysis is to examine gene expression pattern quantitatively and qualitatively in specific cells and tissues. While RefSeq provided by NCBI is utilized as highly accurate reference database for sequence analyses, the reference database for gene expression analyses is not maintained until now. RefEx (Reference Expression dataset; http://togoexp.dbcls.jp/RefEx/) is a challenge to achieve such reference of mammalian gene expression data by different types of methods (EST, GeneChip, iAFLP, and CAGE). Gene expression data by RNA-seq has been rapidly accumulated, and the functional annotation of such sequence from metadata is required and the curation of such data is crucial for the biological interpretation of a bunch of sequences obtained.

In addition to bulk download of the data, RefEx can also be accessed via web interface, and it contains the form in which users can search by gene names, various types of IDs, chromosomal regions in genetic maps, gene family by Interpro, gene expression patterns, and biological categories based on the Gene Ontology (GO).

We will present current status of the project and discuss how to improve the biological curation of transcriptomes.

Semantic Encoding of Complex Information

Steven Vercruysse, Martin Kuiper

Systems Biology group, NTNU (Norwegian University of Science and Technology), Norway

The biosciences are a domain where pieces of information are often highly complex and composite in nature, and where the specific context is often crucial for a correct understanding of a biological system. We took it as a challenge to develop a flexible curation approach that allows a full semantic description of any type of information. Current approaches mainly focus on a table or spreadsheet-based formatting of standardized facts. However, in practice curators often encounter extra relevant context that can not be captured in these predesigned templates. This becomes especially apparent when curating a variety of detailed facts from diverse publications. Also controlled languages only allow partial capture of this variety of facts and context. Therefore we designed a new, general method to capture composite information into a flexible, yet semantically clear format.

We present a flexible method to graphically indicate the structure in any linear series of controlled terms (linked to semantic identifiers). For this, we go beyond representing information with only triples (like RDF); we describe a method to generate structured tuples of arbitrary complexity and depth. We illustrate how most linguistic structures can be built from only three simple connector types (with one being the most prevalent), and how curators can use these 'Visual Syntax Markup' (VSM) connectors to visually indicate the implied structure of any curated sentence. Our approach makes it possible to encode information that is as complex as what natural language presents. We describe the VSM method, its semantics, and its possible application in a user interface. We furthermore show how it may serve as a basis that unites a diverse set of curation initiatives and input forms.
Massively parallel sequencers become widespread and produce unprecedented amounts of sequence reads in many biological fields. DNA Data Bank of Japan (DDBJ) has constructed the international sequence database collaboration (INSDC) together with EBI and NCBI. In 2008, DDBJ has established the DDBJ Read Archive (DRA) to archive raw output data from the new sequencing platforms. DRA archives and provides the raw data sets together with the other two INSDC partners the Sequence Read Archive (SRA) at NCBI and the European Sequence Read Archive (ERA) at EBI.

These new sequencing platforms are also used to count DNA/RNA molecules instead of microarray experiment because of their higher accuracy. Since 2004, DDBJ has operated CIBEX as a repository database for the microarray data. In 2009, we decided to establish a new archive DDBJ Omics Archive (DOR) to efficiently accommodate the massive amounts of quantitative data. DOR integrates array-based CIBEX data. DOR accepts submissions of functional genomics data both from the array- and sequencing-based platforms in collaboration with EBI ArrayExpress. DOR uses the same standards with those of ArrayExpress, namely, MAGE-TAB file format for metadata, MIAME and MINSEQE guidelines for submissions. Thus, the data sets released from DOR are seamlessly exported to ArrayExpress. Moreover, entrances of the submission is unified between the DOR and DRA, in which the submitters once deposit their raw and processed data with necessary metadata, their data will be registered to both databases.

DDBJ continues to serve the biological science with the primary archive databases of DDBJ, DRA and DOR.

PubMedScan: A Keyword-free Recommender of PubMed Articles

In searching for papers in PubMed, most users input keywords or single published papers to find relevant articles. The existing paper search/recommendation systems need appropriate keywords or allow only single abstract as input. We developed a keyword-free PubMed search system, PubMedScan, in which a user specifies the search conditions by multiple papers that the user is interested in. PubMedScan reports newly published articles related to the specified papers by E-mail daily. The relationship of the articles is determined by the related link service in NCBI. We provide two versions, one for local installation and the other for a Web tool. A user inputs PubMed IDs (PMIDs) of the papers of the user’s interest. For easy preparation for the PMIDs, we developed software by which PMIDs can be obtained from PDF or text files of papers under users’ hand. A user can manage multiple topics and register corresponding input papers. The system is useful to collect the proper papers from the flood of new research papers, and is open to the public freely through the URL http://medals.jp/pubmedscan/.

PubMedScan: A Keyword-free Recommender of PubMed Articles

Katsuhiko Murakami\textsuperscript{1\textdagger,3}, Makoto Ogawa\textsuperscript{1,2}, Tadashi Imanishi\textsuperscript{3} and Takashi Gojobori\textsuperscript{3}

\textsuperscript{1}JBIC, Japan
\textsuperscript{2}DYNACOM Co., Ltd., Japan
\textsuperscript{3}BIRC, AIST, Japan

In searching for papers in PubMed, most users input keywords or single published papers to find relevant articles. The existing paper search/recommendation systems need appropriate keywords or allow only single abstract as input. We developed a keyword-free PubMed search system, PubMedScan, in which a user specifies the search conditions by multiple papers that the user is interested in. PubMedScan reports newly published articles related to the specified papers by E-mail daily. The relationship of the articles is determined by the related link service in NCBI. We provide two versions, one for local installation and the other for a Web tool. A user inputs PubMed IDs (PMIDs) of the papers of the user’s interest. For easy preparation for the PMIDs, we developed software by which PMIDs can be obtained from PDF or text files of papers under users’ hand. A user can manage multiple topics and register corresponding input papers. The system is useful to collect the proper papers from the flood of new research papers, and is open to the public freely through the URL http://medals.jp/pubmedscan/.
UniRule - Automatic Annotation In UniProtKB

Alan Bridge¹, Diego Poggioli², Claire O'Donovan², and the UniProt consortium¹,²,³

¹ Swiss Institute of Bioinformatics, Centre Medicale Universitaire,  
  1 rue Michel Servet, CH1211 Geneva 4, Switzerland  
² EMBL European Bioinformatics Institute, Wellcome Trust  
  Genome Campus, Hinxton, Cambridge CB10 2SD, UK  
³ Protein Information Resource, Georgetown University Medical Center,  
  3300 Whitehaven St. NW, Suite 1200, Washington, DC 20007, USA

The UniProt KnowledgeBase (UniProtKB) provides a stable, comprehensive, freely accessible, centralized resource on protein sequences and functional annotation. UniProtKB consists of two sections: UniProtKB/SwissProt, a section containing records that are manually annotated with information extracted from the literature and curator-evaluated computational analysis, and UniProtKB/TrEMBL, a section containing computationally analysed records enriched with automatic annotation and classification. Automatic annotation is an essential complement to manual annotation, which cannot keep pace with either current or projected rates of growth of UniProtKB. To this end UniProt is developing an integrated annotation system termed UniRule, which is based on both manually curated and automatically generated annotation rules and signatures. UniRule provides a variety of annotation types including protein names, general functional annotation in the form of free text and controlled vocabularies, sequence annotation including domains and residues of functional importance, and inferred family relationships. At the time of writing (UniProtKB release 2010_08), UniRule provides annotations for around 35% of more than 11 million UniProtKB/TrEMBL entries. The UniRule system and plans for its future development will be presented.

Rebuilding the Card Catalog – Resolving the Problem of Orphan Enzyme Activities

Shearer, AG¹, Begley, TP², Karp, PD¹, Kothari, A¹, Ngo, C¹, Zarafshar, S¹

¹ SRI International, Menlo Park, CA, United States of America  
² Texas A&M University, College Station, TX, United States of America

Sequence data is the thread that ties a century of enzymology to our era of rapid, facile genome sequencing. The value of each newly sequenced genome directly depends on our ability to reliably assign functions to the genes within that genome. Unfortunately, our ability to figure out what genes do has not kept pace with the dramatic rise in sequencing capacity, and researchers consistently run into the wall of being unable to assign functions to a significant portion of the predicted genes in each new genome.

Disconcertingly, a parallel gaps exists in sequence data for many well-characterized enzymes. Several years ago, we realized that approximately 30% of E.C. enzyme activities (roughly 1,100) had no associated sequence data in any major database. We subsequently carried out a systematic effort to evaluate each putative “orphan” activity with the goal of either finding “lost” sequence data or paving the way for a systematic effort to generate sequence data for those enzymes that were genuinely never sequenced. As we approach the end of the “literature archaeology” phase of this project, we can now report that approximately 200 activities have had their associated sequences unearthed, and that we are preparing to release a comprehensive guide to help the community identify sequences for the genuine orphan activities that remain.
An Automatic Method for Large-scale Functional Annotations Based on a Tree Alignment Algorithm

Masumi Itoh, Kanako O. Koyanagi, Hiroki Arimura, Hidemi Watanabe

Graduate School of Information Science and Technology, Hokkaido University, Japan

Rapid growth of sequencing technology has created the need for fast and accurate functional assignment for large sets of genes. Newly sequenced genes are usually annotated computationally by identifying their orthology relationships to known genes. This approach will not, however, work well in the situation where the number of newly sequenced genomes increases exponentially as in reality, since the ratio of the reference sequences will decrease in the entire sequences analyzed. Although detailed phylogenetic relationships must be needed for annotation of gene sets including paralogs, it is impractical to manually reconstruct and analyze phylogeny of a genome-scale data set. To overcome this problem, we developed a novel automated functional annotation method for a large set of genes. Our method efficiently annotates gene functions by utilizing evolutionary information embedded in functional classification of known genes. Gene functions are hierarchically classified according to their functional similarity, catalytic specificity, localization, expression stage and so on. Because diversification of gene functions is closely related with sequence evolution, the structure of functional classification reflects phylogenetic relationships among genes. The functional classifications are represented as tree structures. In our method, phylogenetic trees of queries are aligned to functional classification trees by a tree alignment algorithm. According to the alignment, query genes are efficiently positioned on functional classification trees, and annotations are transferred from functional classification trees to phylogenetic trees of query genes. Consequently, we can estimate evolution of query genes during functional diversification without reconstructing a large phylogenetic tree including reference sequences. With this approach, we can automatically annotate species-specific paralogs at possible resolution even if their orthologs are absent in function databases in large scale.

HEAT: a New Tool for Gene Set Enrichment Analysis Using Comprehensive Annotation of Human Genes in H-InvDB

Tadashi Imanishi¹, Akiko Ogura Noda¹, and Miho Sera¹

¹BIRC, AIST, Japan

H-InvDB Enrichment Analysis Tool (HEAT) is a new data-mining tool for gene set enrichment analysis based on comprehensive annotations of human genes in H-InvDB. HEAT searches for H-InvDB annotations that are significantly enriched in a user-defined gene set, as compared with the entire H-InvDB representative transcripts. The advantage of HEAT is the wide variety of annotation items used for its analysis: chromosomal bands, InterPro functional domains, Gene Ontology terms, KEGG pathways, H-InvDB gene families/groups, SCOP structural domains, subcellular localization predicted by using the Wolf-PSORT program, tissue-specific gene expression as defined in the H-ANGEL database, and transcription factor binding sites in promoter regions based on JASPAR. HEAT accepts lists of human gene identifiers (IDs) including HUGO gene symbols, accession numbers of INSD (DDBJ/EMBL/GenBank), UniProt accession numbers, Gene IDs, Ensembl Gene IDs, H-InvDB Transcript IDs (HIT) and Locus IDs (HIX), etc. Then, HEAT converts the accepted IDs into HIX using the ID Converter System (http://biodb.jp/), collects various annotations of H-InvDB representative transcripts, and conducts statistical tests by using Fisher's exact probability. The output of HEAT is a simple report of annotations commonly found among the query genes, which is very useful to grasp the property of a particular gene set. HEAT is freely available at http://hinv.jp/HEAT/.
Human Transcript Database Search Showed Existence Of Extremely Short Introns

Makoto K. SHIMADA\(^1\), Noriko HARAGUCHI\(^1\), Akila MAYEDA\(^1\)

\(^1\)Institute for Comprehensive Medical Science, Fujita Health University, Japan

Since pre-mRNA splicing is processed by a spliceosome that is a huge complex consisting of RNA and as many as 200 proteins, it is considered that minimum intron size is limited by molecular size of the spliceosome. To understand splicing mechanism, we investigate human intron size distribution using annotated transcriptome database, H-InvDB. Distribution of obtained intron length shows that there is a mode at 83 nt in length with 4049 transcripts and number of introns decreases drastically in shorter than 65 nt, where numbers of transcripts in each locus also decrease. However, some introns less than 65 nt are observed, both shows high Codon Adaptation Index (CAI) and also observed orthologous transcripts in other mammals. These suggest that there is a limitation in minimum length to spliced out in general splicing mechanism of human, but also unknown mechanism may be to splice small intron exceeding the limitation.

The Latest Information and Perspective of Human Gene and Protein Database (HGPD) as Human Proteome Study

Yukio MARUYAMA\(^1\), Yoshifumi KAWAMURA\(^1\), Takao ISOGAI\(^2\), Nobuo NOMURA\(^3\) and Naoki GOSHIMA\(^3\)

\(^1\)Japan Biological Informatics Consortium, JAPAN
\(^2\)Graduate School of Pharmaceutical Sciences, The University of Tokyo, JAPAN
\(^3\)National Institute of Advanced Industrial Science and Technology, JAPAN

HGPD (http://www.HGPD.jp/) is a unique database that stores information on a set of human Gateway entry clones in addition to protein expression data. HGPD was launched in November 2008 and, since then, 33,275 human Gateway entry clones have been constructed from ORFs of full-length cDNAs, thus representing the largest collection in the world. Sequence information and protein expression of Gateway entry clones can also be retrieved from HGPD. The majority of analysis data for cDNA sequences in HGPD are shared with the FLJ Human cDNA Database (http://flj.hinv.jp/v01/cgi/index.cgi), which was constructed as human cDNA sequence analysis database focusing on mRNA varieties resulting from variations in transcription start site (TSS) and splicing.

In this update, we constructed an additional 9,974 Gateway entry clones, giving a total of 43,249, corresponding to approximately 2,000 loci. Sequence summary, nucleotide and amino acid sequence data for these clones (v 2.0) can be obtained from download site. We named this human Gateway entry clone resource “HUPEX” (Human Proteome Expression-resource). And, we expressed 4,457 human proteins with a C-terminal V5 and His tag and analyzed them using SDS-PAGE, giving a total of 17,821. In the near future, we will construct 11,000 image data of subcellular localization of the proteins with fluorescent proteins at the N- or C-terminal which are expressed in HeLa cells by using Gateway entry clones. And, we will link ESPRESSO (http://mbs.cbrce.jp/ESPRESSO/) using ORF sequences of the entry clones, allowing us to acquire prediction of protein expression and solubility results. Moreover, we will redesign the web interface.
The Bioinformatics Links Directory: A Community Curated Collection of Bioinformatics Links, Tools and Databases

David Yim, Michelle D. Brazas, Joseph Yamada, Raffi Melkon, Winston Yeung and B.F. Francis Ouellette

Ontario Institute for Cancer Research, Toronto, Ontario, Canada

The Bioinformatics Links Directory (http://bioinformatics.ca/links_directory/) is a compendium of useful bioinformatics links organized in an intuitive, research-task hierarchy. This directory was first initiated in 2002, and has been maintained by a number of people working with the Ouellette Laboratory over the last 8 years. Many of the links in the Links Directory come from the annual NAR Web Server issue, with significant input and suggestions from the user community. More than 80% of all links have one or more PubMed reference associated, providing detailed information about the resource. With the recent migration to the Drupal 6 content management system, a number of new features have been introduced. These include:

- An upgraded Links Directory that now includes tools and databases in addition to links;
- Better content management and the introduction of metadata to allow for more meaningful content searches;
- Ability to facilitate and monitor community input and directory upkeep;
- Introduction of tags from user input and PubMed MeSH terms for better content curation;
- Development of user groups to encourage collaborative content curation and maintenance within a group or an institution;
- Web Services/API to receive submissions from journal publishers for content diversity.

The upgraded Links Directory will provide and allow for a better user experience, resulting in faster identification of required bioinformatics’ resources. Continued community input to supplement links with supporting documentation and educational materials, will ensure that the Links Directory remains an invaluable bioinformatics resource for everyone.
Exhibition

Exhibitions are open during the following hours:
October 12 (Tue), 12:00-17:30
October 13 (Wed), 8:30-17:30
October 14 (Thu), 8:30-12:30

<table>
<thead>
<tr>
<th>No.</th>
<th>EXHIBITORS</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>JAPAN BIOLOGICAL INFORMATICS CONSORTIUM</td>
</tr>
<tr>
<td>2</td>
<td>Integrated Database and Systems Biology Team, BIRC, AIST</td>
</tr>
<tr>
<td>3</td>
<td>Oxford University Press</td>
</tr>
<tr>
<td>4</td>
<td>FUJITSU LIMITED</td>
</tr>
<tr>
<td>5</td>
<td>Database Center for Life Science, ROIS</td>
</tr>
<tr>
<td>6</td>
<td>RIKEN Integrated Databases</td>
</tr>
<tr>
<td>Name</td>
<td>Organization</td>
</tr>
<tr>
<td>-----------------------------</td>
<td>-------------------------------------------------------------------------------</td>
</tr>
<tr>
<td>Santos Alberto</td>
<td>Center for Genomic Sciences</td>
</tr>
<tr>
<td>Vigilante Alessandra</td>
<td>University of Naples Federico II</td>
</tr>
<tr>
<td>Kiyoko F. Aoki-Kinoshita</td>
<td>Soka University</td>
</tr>
<tr>
<td>Amos Bairoch</td>
<td>Swiss Institute of Bioinformatics and University of Geneva</td>
</tr>
<tr>
<td>Ana Barat</td>
<td>Dublin City University (DCU)</td>
</tr>
<tr>
<td>Alex Bateman</td>
<td>Wellcome Trust Sanger Institute</td>
</tr>
<tr>
<td>Tanya Z. Berardini</td>
<td>Carnegie Institution for Science</td>
</tr>
<tr>
<td>Hidemasa Bono</td>
<td>Research Organization of Information and Systems</td>
</tr>
<tr>
<td>Alan J. Bridge</td>
<td>Swiss Institute of Bioinformatics</td>
</tr>
<tr>
<td>Sarah W. Burge</td>
<td>EMBL-EBI Wellcome Trust Genome Campus</td>
</tr>
<tr>
<td>Tzen-Yuh Chiang</td>
<td>National Cheng Kung Univ.</td>
</tr>
<tr>
<td>Maria Costanzo</td>
<td>Stanford University School of Medicine</td>
</tr>
<tr>
<td>Delphine Dauga</td>
<td>IBDM-CNRS</td>
</tr>
<tr>
<td>Maguire Eamonn</td>
<td>University of Oxford</td>
</tr>
<tr>
<td>Toshinori Endo</td>
<td>Hokkaido University</td>
</tr>
<tr>
<td>Marcus Ennis</td>
<td>EMBL-EBI Wellcome Trust Genome Campus</td>
</tr>
<tr>
<td>Yasuyuki Fujii</td>
<td>Okayama University</td>
</tr>
<tr>
<td>Takatomo Fujisawa</td>
<td>Kazusa DNA Research Institute</td>
</tr>
<tr>
<td>Nobuyuki Fujita</td>
<td>National Institute of Technology and Evaluation</td>
</tr>
<tr>
<td>Socorro Gama-castro</td>
<td>Center for Genomic Sciences</td>
</tr>
<tr>
<td>Michael J. Gardner</td>
<td>EMBL-EBI Wellcome Trust Genome Campus</td>
</tr>
<tr>
<td>George Garrity</td>
<td>Michigan State University</td>
</tr>
<tr>
<td>Pascale Gaudet</td>
<td>Northwestern University</td>
</tr>
<tr>
<td>Takashi Gojobori</td>
<td>National Institute of Genetics</td>
</tr>
<tr>
<td>Susumu Goto</td>
<td>Kyoto University</td>
</tr>
<tr>
<td>Yuichiro Hara</td>
<td>BIRC, AIST</td>
</tr>
<tr>
<td>Winston Hide</td>
<td>Harvard School of Public Health</td>
</tr>
<tr>
<td>Yuichiro Hira</td>
<td>Keio University</td>
</tr>
<tr>
<td>Kohji Hotta</td>
<td>Keio University</td>
</tr>
<tr>
<td>Name</td>
<td>Institution</td>
</tr>
<tr>
<td>-----------------------------</td>
<td>-----------------------------------------------------------------------------</td>
</tr>
<tr>
<td>Tzu-Ting Hsu</td>
<td>National Central University</td>
</tr>
<tr>
<td>Chi-Chun Huang</td>
<td>National Cheng Kung University</td>
</tr>
<tr>
<td>Natsuko Ichikawa</td>
<td>National Institute of Technology and Evaluation</td>
</tr>
<tr>
<td>Kazuho Ikeo</td>
<td>National Institute of Genetics</td>
</tr>
<tr>
<td>Tadashi Imanishi</td>
<td>BIRC, AIST</td>
</tr>
<tr>
<td>Takeshi Itoh</td>
<td>National Institute of Agrobiological Sciences</td>
</tr>
<tr>
<td>Masumi Itoh</td>
<td>Graduate School of Information Science, Hokkaido University</td>
</tr>
<tr>
<td>Young Jasmine</td>
<td>Protein Data Bank</td>
</tr>
<tr>
<td>Kwangjoon Jeong</td>
<td>Dept. of Microbiology, Chonnam National University Medical School</td>
</tr>
<tr>
<td>Chris Jordan</td>
<td>Texas Advanced Computing Center</td>
</tr>
<tr>
<td>Zara Josephs</td>
<td>EMBL-EBI Wellcome Trust Genome Campus</td>
</tr>
<tr>
<td>Eli Kaminuma</td>
<td>National Institute of Genetics</td>
</tr>
<tr>
<td>Minoru Kanehisa</td>
<td>Kyoto University, Bioinformatics Center</td>
</tr>
<tr>
<td>Renate Kania</td>
<td>Heidelberg Institute for Theoretical Studies</td>
</tr>
<tr>
<td>Toshiaki Katayama</td>
<td>Univ Tokyo</td>
</tr>
<tr>
<td>Yoshihiro Kawahara</td>
<td>NIAS</td>
</tr>
<tr>
<td>Hideya Kawaji</td>
<td>RIKEN Omics Science Center</td>
</tr>
<tr>
<td>Prasad Keshava</td>
<td>Institute of Bioinformatics</td>
</tr>
<tr>
<td>Akira Kinjo</td>
<td>Institute for Protein Research, Osaka University</td>
</tr>
<tr>
<td>Akihiko Kishimoto</td>
<td>Japan Biological Informatics Consortium</td>
</tr>
<tr>
<td>Norio Kobayashi</td>
<td>Bioinformatics And Systems Engineering (BASE) division, RIKEN, Japan</td>
</tr>
<tr>
<td>Yuichi Kodama</td>
<td>Center for Information Biology and DNA Data Bank of Japan, National Institute of Genetics, Japan</td>
</tr>
<tr>
<td>Chai Yin Kok</td>
<td>Wellcome Trust Sanger Institute</td>
</tr>
<tr>
<td>Yusuke Komiyama</td>
<td>Kyoto University</td>
</tr>
<tr>
<td>Takehide Kosuge</td>
<td>National Institute of Genetics</td>
</tr>
<tr>
<td>Lydie Lane</td>
<td>Swiss Institute of Bioinformatics</td>
</tr>
<tr>
<td>Hughes Laron</td>
<td>Northwestern University</td>
</tr>
<tr>
<td>Stan Laulederkind</td>
<td>Medical College of Wisconsin</td>
</tr>
<tr>
<td>Sung Shin Lee</td>
<td>National Institute of Agrobiological Sciences</td>
</tr>
<tr>
<td>Joo-Young Lee</td>
<td>Clinical vaccine R&amp;D Center</td>
</tr>
<tr>
<td>Yixue Li</td>
<td>Shanghai Center for Bioinformation Technology</td>
</tr>
<tr>
<td>Damien Lieberherr</td>
<td>UniProtKB/Swiss-Prot</td>
</tr>
<tr>
<td>Name</td>
<td>Institution</td>
</tr>
<tr>
<td>-------------------------------</td>
<td>--------------------------------------------------</td>
</tr>
<tr>
<td>Frederique Lisacek</td>
<td>Swiss Institute of Bioinformatics</td>
</tr>
<tr>
<td>Michael Livstone</td>
<td>Princeton University</td>
</tr>
<tr>
<td>Juncai Ma</td>
<td>Institute of Microbiology, CAS</td>
</tr>
<tr>
<td>Yuko Makita</td>
<td>RIKEN</td>
</tr>
<tr>
<td>Kentarou Mamiya</td>
<td>Japan Biological Informatics Consortium</td>
</tr>
<tr>
<td>Aron Marchler-Bauer</td>
<td>National Library of Medicine</td>
</tr>
<tr>
<td>Akihiro Matsuya</td>
<td>Japan Biological Informatics Consortium</td>
</tr>
<tr>
<td>Ramil P. Mauleon</td>
<td>International Rice Research Institute</td>
</tr>
<tr>
<td>Raja Mazumder</td>
<td>PIR, Georgetown University Medical Center</td>
</tr>
<tr>
<td>Fiona M. McCarthy</td>
<td>Mississippi State University</td>
</tr>
<tr>
<td>Lijowski Michal</td>
<td>Washington University in St. Louis</td>
</tr>
<tr>
<td>Magrane Michele</td>
<td>EMBL- EBI</td>
</tr>
<tr>
<td>Ilene Mizrachi</td>
<td>National Center for Biotechnology Information, NLM, NIH</td>
</tr>
<tr>
<td>Barend Mons</td>
<td>Netherlands Bioinformatics Centre</td>
</tr>
<tr>
<td>Monica C. Munoz-Torres</td>
<td>Georgetown University</td>
</tr>
<tr>
<td>Katsuhiko Murakami</td>
<td>Japan Biological Informatics Consortium</td>
</tr>
<tr>
<td>Kenta Nakai</td>
<td>University of Tokyo</td>
</tr>
<tr>
<td>Mitsuteru Nakao</td>
<td>DBCLS</td>
</tr>
<tr>
<td>Rajni Nigam</td>
<td>Medical College of Wisconsin</td>
</tr>
<tr>
<td>Takaaki Nishioka</td>
<td>Keio University</td>
</tr>
<tr>
<td>Marc-Alexandre Nolin</td>
<td>Laval University</td>
</tr>
<tr>
<td>Hisataka Numa</td>
<td>National Institute of Agrobiological Sciences</td>
</tr>
<tr>
<td>Claire O'donovan</td>
<td>UniProt, European Bioinformatics Institute</td>
</tr>
<tr>
<td>Akiko Ogura-Noda</td>
<td>BIRC, AIST</td>
</tr>
<tr>
<td>Hajime Ohyanagi</td>
<td>National Institute of Genetics</td>
</tr>
<tr>
<td>Shinobu Okamoto</td>
<td>The Database Center for Life Science</td>
</tr>
<tr>
<td>Hiromasa Ono</td>
<td>Database Center for Life Science</td>
</tr>
<tr>
<td>Sandra Orchard</td>
<td>EMBL- EBI</td>
</tr>
<tr>
<td>Francis Ouellette</td>
<td>Ontario Institute for Cancer Research</td>
</tr>
<tr>
<td>Guvanch Ovezmyradov</td>
<td>Max Planck Institute for Experimental Medicine</td>
</tr>
<tr>
<td>Steve Pettifer</td>
<td>University of Manchester</td>
</tr>
<tr>
<td>Rocca-Serra Philippe</td>
<td>University of Oxford</td>
</tr>
<tr>
<td>Name</td>
<td>Institution/University</td>
</tr>
<tr>
<td>-----------------------------</td>
<td>------------------------------------------------------------</td>
</tr>
<tr>
<td>Deigo Poggioli</td>
<td>EMBL-EBI Wellcome Trust Genome Campus</td>
</tr>
<tr>
<td>Betina M. Porcel</td>
<td>GENOSCOPE, INSTITUT DE GENOMIQUE, CEA</td>
</tr>
<tr>
<td>Sylvain Poux</td>
<td>UniProtKB/Swiss-Prot</td>
</tr>
<tr>
<td>Anuradha Pujar</td>
<td>Boyce Thompson Institute</td>
</tr>
<tr>
<td>Lorna Richardson</td>
<td>EMAGE</td>
</tr>
<tr>
<td>Fabio Rinaldi</td>
<td>University of Zurich</td>
</tr>
<tr>
<td>Ryuichi Sakate</td>
<td>National Institute of Biomedical Innovation (NIBIO)</td>
</tr>
<tr>
<td>Nozomu Sakurai</td>
<td>Kazusa DNA Research Institute</td>
</tr>
<tr>
<td>Mara Sangiovanni</td>
<td>University of Naples Federico II</td>
</tr>
<tr>
<td>Susanna-Assunta Sansone</td>
<td>University of Oxford</td>
</tr>
<tr>
<td>Naoki Sato</td>
<td>University of Tokyo</td>
</tr>
<tr>
<td>Mary L. Schaeffer</td>
<td>USDA-ARS</td>
</tr>
<tr>
<td>Erik Segerdell</td>
<td>Oregon Health &amp; Science University</td>
</tr>
<tr>
<td>Miho Sera</td>
<td>BIRC, AIST</td>
</tr>
<tr>
<td>Venkataraman Shanmugasundaram</td>
<td>MRC Human Genetics Unit</td>
</tr>
<tr>
<td>Alexander Shearer</td>
<td>SRI International</td>
</tr>
<tr>
<td>Kei-ichi Shimazaki</td>
<td>Hokkaido University</td>
</tr>
<tr>
<td>Masayo Shindo</td>
<td></td>
</tr>
<tr>
<td>Rosemary Shrestha</td>
<td>CIMMYT (International Maize and Wheat Improvement Center)</td>
</tr>
<tr>
<td>Yuka Takahashi</td>
<td>National Institute of Genetics</td>
</tr>
<tr>
<td>Jun-ichi Takeda</td>
<td>BIRC, AIST</td>
</tr>
<tr>
<td>Tin-Wee Tan</td>
<td>National University of Singapore</td>
</tr>
<tr>
<td>Tsuyoshi Tanaka</td>
<td>National Institute of Agrobiological Sciences</td>
</tr>
<tr>
<td>Tatiana Tatusova</td>
<td>NIH/NLM/NCBI</td>
</tr>
<tr>
<td>Isaak Y. Tecle</td>
<td>Cornell University</td>
</tr>
<tr>
<td>Petra Ten Hooppen</td>
<td>EMBL-EBI Wellcome Trust Genome Campus</td>
</tr>
<tr>
<td>Anne Thessen</td>
<td>Marine Biological Laboratory</td>
</tr>
<tr>
<td>Paul D. Thomas</td>
<td>University of Southern California</td>
</tr>
<tr>
<td>Dave Thorne</td>
<td>University of Manchester</td>
</tr>
<tr>
<td>Kentaro Tomii</td>
<td>National Institute of Advanced Industrial Science and Technology (AIST)</td>
</tr>
<tr>
<td>Tetsuro Toyoda</td>
<td>RIKEN BASE</td>
</tr>
<tr>
<td>Alessandra Traini</td>
<td>University of Naples Federico II</td>
</tr>
<tr>
<td>Name</td>
<td>Institution</td>
</tr>
<tr>
<td>---------------------</td>
<td>--------------------------------------------------</td>
</tr>
<tr>
<td>Rie Tsuchiya</td>
<td>National Institute of Genetics</td>
</tr>
<tr>
<td>Andrei Turinsky</td>
<td>Hospital for Sick Children</td>
</tr>
<tr>
<td>Keisuke Ueno</td>
<td>Hokkaido University</td>
</tr>
<tr>
<td>Kimberly Van Auken</td>
<td>WormBase, California Institute of Technology</td>
</tr>
<tr>
<td>Ceslovas Venclovas</td>
<td>Inst Biotech</td>
</tr>
<tr>
<td>Steven Vercruysse</td>
<td>NTNU: Norwegian University of Science and Technology</td>
</tr>
<tr>
<td>Xiaodong Wang</td>
<td>WormBase, California Institute of Technology</td>
</tr>
<tr>
<td>Trish Whetzel</td>
<td>National Center for Biomedical Ontology / Stanford University</td>
</tr>
<tr>
<td>Owen White</td>
<td>University of Maryland School of Medicine</td>
</tr>
<tr>
<td>Melanie Wilson</td>
<td>Oregon Health and Science University</td>
</tr>
<tr>
<td>Ulrike Wittig</td>
<td>Heidelberg Institute for Theoretical Studies (HITS)</td>
</tr>
<tr>
<td>Cheng-Chieh Wu</td>
<td>Academia Sinica, Taipei, Taiwan</td>
</tr>
<tr>
<td>Hsin-Hui Wu</td>
<td>Saint Louis University</td>
</tr>
<tr>
<td>Lu Xinghua</td>
<td>Dept of Biomedical Informatics, University of Pittsburgh</td>
</tr>
<tr>
<td>Yasunori Yamamoto</td>
<td>Database Center for Life Science</td>
</tr>
<tr>
<td>Chisato Yamasaki</td>
<td>BIRC, AIST</td>
</tr>
<tr>
<td>Riu Yamashita</td>
<td>Frontier Research Initiative, Institute of Medical Science, University of Tokyo</td>
</tr>
<tr>
<td>Yukiko Yamazaki</td>
<td>National Institute of Genetics</td>
</tr>
<tr>
<td>Toru Yao</td>
<td>BIRC, AIST</td>
</tr>
<tr>
<td>Kouki Yonezawa</td>
<td>Graduate School of Information Science and Technology, Hokkaido University</td>
</tr>
<tr>
<td>Karen Yook</td>
<td>WormBase, Calif Institute of Technology</td>
</tr>
<tr>
<td>Peifen Zhang</td>
<td>Carnegie Institution</td>
</tr>
</tbody>
</table>
Hitachi GP
Bioinformatics Solution

Features of our Bioinformatics Solution

Well-Designed User Interface
- Excellent Usability
- Beautiful Look and Feel

Served by Bioinformatician
- All members have skills in bioinformatics

Solution Menu

Web-based Application Development
- All sorts of bioinformatics browsers and data-bases

Data Analysis
- Sequence Analysis (NGS)
- Statistical Analysis (Expression, GWAS)
- Evolutionary Analysis

Hitachi GP reserves the right to make changes to the services described herein at any time without notice.

Bioinformatics Group, Research & Development Center, Solution Division 4, Hitachi Government & Public Corporation System Engineering, Ltd.
2-4-18 Toyo, Koto-ku, Tokyo 135-8633, Japan
http://www.gp.hitachi.co.jp/service/bioinfo/index.html
実験医学

8月号 (Vol.28 No.13)
Protein Kinesisを解き明かす
オルガナの世界

9月号 (Vol.28 No.14)
神経科学と臨床研究の連携がもたらす
精神疾患への統合的アプローチ

10月号 (Vol.28 No.16)
RNAの時空間的動態をみる

改訂第5版
新 遺伝子工学ハンドブック

がんの分子標的と
治療薬事典

科研費獲得の
方法とコツ

創薬・タンパク質研究のための
プロテオミクス解析

分子生物学講義中経 番外編
生物の多様性と
進化の驚異

バイオ実験
超基本Q&A 改訂版

日々の実験・研究生活に役立つ 羊土社の新刊書籍!!

発行
羊土社

〒101-0052 東京都千代田区神田小川町2-5-1
TEL 03（5282）1211 FAX 03（5282）1212
E-mail : ogyo@yodosh.co.jp
URL : http://www.yodosh.co.jp/

ご注文は最寄りの書店、または小社営業部まで
Image adapted from Chea and Livesay. *BMC Bioinformatics* 2007, 8:153
Do Less, Get More !!

DNASIS® Taxon

All-in-one taxonomic identification suite with powerful in-house sequence database management. Perfect for bacterial/virus identification, environment sampling, and so on!

- Full automatic identification from sequence trimming to taxonomic tree
- Built-in taxonomic database and 16S rDNA database
- Powerful custom sequence database management
- Homology search (BLAST)
- Multiple alignment (ClustalW, MUSCLE)
- Supports in-house and Internet sequence database search
- Interactive identification result viewing with intuitive taxonomic tree browser and report tables.

DNASIS® MAX

Bioinformatics software for DNA/RNA/Amino Acid sequence analysis.

- Edit, annotate and analyze DNA, RNA and amino acid sequences with beautiful Easy-to-Use GUI
- Homology search, multiple alignment, Restriction enzyme and many other analysis algorithms
- Sequence assembly (Phred/Phrap, optional)
- Automatic analysis workflows (DNASpace, optional)

MasterPlex® ReaderFit

Software dedicated to Microplate Reader Data Analysis. Get Fast Accurate Results Like Never Before. Perfect for ELISA and all quantitative assays.

- Import CSV data file from most of common Microplate readers
- Automatic 4PL, SPL curve fitting optimization with weighting
- Multi-plate and Multi-plex capability
- Publication quality charts and robust report generator
- Automatic EC50/IC50 Determination
- Built-in Quality Control with Automatic Outlier Detection
- FDA 21 CFR Part 11, GxP (GCP/GLP/GMP) Compliance (Optional)

NOTE: Sold as “DNASIS Pro” in Japan.

Hitachi Solutions

Japan
Hitachi Solutions, Ltd. (Life Science Division)
1-1-49, Sasebo-chō, Tsukumi-ku, Yokohama, 223-0045
Tel: +81-45-500-5111 Fax: +81-45-500-5124
Email: info@dnas.jp
Web: http://www.dnasis.jp/

Worldwide
Hitachi Solutions America, Inc. (MiraBio Group)
601 Gateway Blvd. Suite 100, South San Francisco, CA 94080, USA
Tel: +1-650-615-7600 Fax: +1-650-615-7639
Email: info@mira.bio
Web: http://www.mira.bio.com/
Tokyo Academic Park / Access map

■ Odaiba area map

■ Vicinity map

2-2-1, Aomi, Koto-ku, Tokyo 135-8630
TEL 03-5520-6055

■ Access by train

东京研究交流大学村
TOKYO ACADEMIC PARK