
Program Changes

i) An additional talk [S1-5, Session 1]

10:30-12:30, Oct. 12th, Tuesday

Session 1: Large Scale, Automated, and Predictive Annotations

S1-1/10:30 *Comparison of Computationally- and Manually-Assigned Gene Ontology Annotations to Improve Functional Characterization of Gene Products*

S1-2/10:55* *Automatic Protein Clustering as a Basis of Automatic Annotation*

S1-3/11:20* *Transcriptome in a Dynamic System with Next Gen Sequencers*

S1-4/11:45* *UniRule - Automatic Annotation In UniProtKB*

S1-5/12:10 *Towards Developing Common Standards for Genome Sequence and Annotation*

Tatiana Tatusova (NCBI, NLM, NIH, USA)

The abstract for S1-5 by Tatiana Tatusova is found in P9-4 (poster session), and the poster P9-4 has been canceled instead.

*Starting times have been changed.

ii) Additional posters

P15-10 *Microbial Genome Annotation Pipeline (MiGAP)*

Hideaki Sugawara

P15-11 *TargetMine, an Integrated Data Warehouse for Candidate Gene Prioritisation*

Lokesh P. Tripathi

P15-12 *Development of Thermodynamic Databases of Biomolecules and Their Interactions*

Shaji Kumar

iii) Abstracts for additional posters

P15-10

Microbial Genome Annotation Pipeline (MiGAP)

Hideaki Sugawara¹, Akira Ohyama², Hiroshi Mori³, Ken Kurokawa³

¹National Institute of Genetics, 1111 Yata, Mishima, Shizuoka 411-8540, Japan

²*in silico* biology, Inc., SOHO Station 706, 24-8 Yamashita-cho, Naka-ku, Yokohama, 231-0023 Japan

³Dept. of Biological Information, TITECH, Midori-ku, Yokohama, 226-8501, Japan

More than 1,000 microbial complete genomes have been sequenced as of September 2010 and the rate of sequencing will rocket ahead thanks to the 2nd and 3rd generation sequencers. However, the tsunami of sequence data does not necessarily mean the increase of our knowledge on microbes. The sequences have to be annotated. Microbial Genome Annotation Pipeline (MiGAP) provides novice and old pro alike with a mechanical annotation to microbial contigs and genomes.

MiGAP identifies ORFs and RNA regions and infers the functions of ORFs by referring to highly evaluated public databases. MiGAP has the following three modes of the operation:

- b-MiGAP provides analysis by the default setting of programs, parameters and the reference databases. The user is required to just give sequences to MiGAP to get the annotation
- s-MiGAP provides the user with the freedom to select programs, parameters and the reference databases.
- g-MiGAP provides the user with the function of add his/her own tools and databases to the pipeline in addition to s-MiGAP function.

MiGAP has been open to the public since June 2008 and processed monthly, on average, about 58 mega base pairs of 40 jobs to predict 64,000 CDSs.

P15-11

TargetMine, an integrated data warehouse for candidate gene prioritisation

Yi-An Chen¹, Lokesh P. Tripathi¹, Kenji Mizuguchi¹

¹*National Institute of Biomedical Innovation, 7-6-8 Asagi Saito Ibaraki-City Osaka 567-0085, Japan*

Prioritising candidate genes for further experimental characterisation is a non-trivial challenge in biomedical research. An integrated approach that combines results from multiple data types is best suited for optimal target discovery. We have developed TargetMine, an integrated data warehouse for selection of target genes and proteins for experimental characterisation and drug discovery. TargetMine utilises the flexible InterMine framework, which ensures that different types of biological data can be readily added and analysed to generate new hypothesis. It enables complicated searches that are difficult to perform with existing tools and thus assists in efficient target prioritisation. We propose an objective protocol for identification of candidate genes for further investigations and show that TargetMine has been effectively employed for target selection in the investigation of Hepatitis C virus (HCV) pathogenesis.

P15-12

Development of Thermodynamic Databases of Biomolecules and Their Interactions

Shaji Kumar, Akinori Sarai

*Department of Bioscience and Bioinformatics,
Kyushu Institute of Technology, Fukuoka, Japan*

ProTherm and ProNIT [1-2] are two thermodynamic databases that contain experimentally determined thermodynamic parameters of protein stability and protein–nucleic acid interactions, respectively. Correlation of the thermodynamic data available in these databases and their relationship with the structural parameters would help in elucidating biological processes and also be useful in the process of drug design. Researchers need to query across multiple autonomous and heterogeneous data sets in order to achieve this to facilitate good decision making and the right approach for a specific problem. Integration of heterogeneous data from different databases has become a central problem of bioinformatics field. ProTherm and ProNIT are now part of a centralized database integration project at Database Center for Life Science (DBCLS), to facilitate the analysis and data processing easier for researchers who need to correlate data from different sources. In order to facilitate the data integration, we provide an XML format to distribute the thermodynamic data. A proper XML schema has been designed to define the building blocks of the XML document and also to define the document structure of the ProNIT XML data. A common semantics is an essential element in the database integration. So, we are developing an ontology for thermodynamic data of biomolecules and interactions to facilitate the data integration. So far, we collect the thermodynamic data from the literature. The data collection and extraction from printed articles are very time-consuming. Thus, we are trying to ease the process by using the text-mining technology, in collaboration with DBCLS.

[1] M.M. Gromiha, J. An, H. Kono, M. Oobatake, H. Uedaira, A. Sarai, *Nucleic Acids Res.*, 27, 286–288 (1999).

[2] M. D. Shaji Kumar, K. A. Bava, M. M. Gromiha, P. Prabakaran, K. Kitajima, H. Uedaira and A. Sarai, *Nucleic Acids Res.*, 34, D204-206 (2006).