

## O-7

**EVOLUTIONARY-BASED BIOINFORMATICS ANALYSIS OF PRESYNAPTIC GENES**

D D. Hadley<sup>1,2</sup>, T K Murphy<sup>2</sup>, O Valladares<sup>2</sup>, L Ungar<sup>1,4</sup>, J Kim<sup>1,4,5</sup>, M Bucan<sup>1,2,3</sup>

Penn Center for Bioinformatics / Genomics and Computational Biology Graduate Group<sup>1</sup>, Department of Genetics<sup>2</sup>, School of Medicine<sup>3</sup>, Department of Computer & Information Sciences / School of Engineering and Applied Sciences<sup>4</sup>, Department of Biology / School of Applied Sciences<sup>5</sup>; University of Pennsylvania, Philadelphia PA 19104

To facilitate identification of *cis*-regulatory elements involved in the transcriptional and translational control of gene expression in the neuronal synapse, we initiated a large-scale comparative analysis of genes implicated in presynaptic function. Although annotation of both protein- and non-protein-coding annotation is available through a number of public databases, such datasets are highly automated and somewhat limited in resolution. Thus, we sought to complement these genome-wide efforts by focusing on 130 presynaptic genes (63Mb), and providing highly curated, in-depth annotation of their genomic neighborhoods. Evolutionary analysis combined with bioinformatics approaches, represent a powerful mechanism for understanding the genomic landscape, and we have employed such methods here. By first focusing on regions undergoing purifying selection and then determining various measures of biological importance *in silico*, we prioritize genomic elements for both *in vitro* and *in vivo* verification. In particular, computational approaches include determining the rate of protein evolution, estimating codon bias on coding sequences, and calculating the folding energy of noncoding elements. In this paradigm, we have annotated novel transcripts, missed exons, candidate miRNAs and putative miRNA targets. In addition to considering genes individually, we also consider them in the context of their gene family where applicable. In so doing, we are beginning to explain diverse gene ontologies, gene expressions and other readily available phenotypes from an evolutionary perspective.

## Bioinformatics

Oral Presentation

Saturday November 5

4.45pm – 5.00pm

O-8

**INTELLIGENT INFERENCES IN THE OMIC SPACE: TOWARD HIGH-THROUGHPUT *IN SILICO* POSITIONAL CLONING**

N Kobayashi, Y Mochizuki, Y Hasegawa, N Heida, K Player, T Toyoda  
Genomic Sciences Center, RIKEN, Yokohama, Japan

Since whole genome sequences were first elucidated, knowledge-based ranking of candidate genes has become one of the most important bioinformatics tasks in the forward-genetics and positional-cloning approaches to identify phenotype-responsible gene mutations. This task requires creating a form of artificial intelligence that solves a genetics researcher's problem by learning computationally a vast amount of information elucidated from data ranging from genomic to phenomic levels.

In order to integrate various *omic* annotations and interactions in the databases, we have been applying the coordinate-based integration methodology that we proposed previously [1], rather than the conventional identifier-based integration. The advantage of coordinate-based integration is that it can unite different types of data items having complex many-to-many correspondences that are difficult to represent by identifier-based referencing. Our method is especially effective for integration of distributed databases, since we can now share common genomic sequences to realize the integration of world-wide distributed data based on consistent positions on the same coordinate axes.

By adopting our methodology, we have developed a system that suggests highly promising candidate genes in a given chromosomal interval. The system employs a full-text search engine against biological literature (the Medline abstracts) and handles other *omic* knowledge accumulated in our databases. In the case where few candidates are found by direct keyword search, the system automatically proceeds to infer other candidates through biological networks. Another improvement is that we have added original contextual words to the gene-name dictionary so as to improve the accuracy of gene-name recognition from the literature.

[1] Toyoda, T. and Wada, A. (2004) *Bioinformatics* 20, 1759-65.

**Bioinformatics**

**Oral Presentation**

**Saturday November 5**

**5.00pm – 5.15pm**

**O-9**

**EMAGE – A SPATIAL DATABASE OF GENE EXPRESSION PATTERNS IN THE DEVELOPING MOUSE EMBRYO. TOWARDS A TOOL FOR COMPUTATIONAL IDENTIFICATION OF SETS OF CO-EXPRESSED GENES FROM IN SITU EXPERIMENTS**

J Christiansen, L Richardson, S Venkataraman, P Stevenson, N Burton, Y Yang, C Semple, R Baldock, D Davidson  
MRC Human Genetics Unit, Edinburgh, United Kingdom

EMAGE is a database of spatially mapped in situ gene expression patterns in the developing mouse embryo.

All EMAGE data is housed in a standard framework: the EMAP Digital Atlas of Mouse Development. This consists of at least one representative 3D digital embryo model at most Theiler stages (TS) as well as a standardised nomenclature for the anatomical structures that are present at every TS of development. The digital embryo models are 3D objects and virtual sections can be cut in any plane to reveal internal anatomical detail.

Raw incoming data images are converted to digital representations and mapped spatially into the corresponding regions within the embryo models. At TS07-14 the 3D standard models have anatomical regions defined within them and data spatially mapped into these models is automatically annotated to the corresponding text terms for these structures. This is accompanied by further manual text annotation.

Current data searching can be done spatially by defining a region of interest in a particular embryo model, or by using text terms to find genes expressed within the region and/or named structure. We show in pilot studies that the database can also be used to identify groups of co-expressed genes by hierarchical clustering and domain intersection analyses.

Free software to search EMAGE can be downloaded from <http://genex.hgu.mrc.ac.uk>. The same software can also be used to prepare private databases for in-lab data management or to prepare electronic submissions to EMAGE. Alternatively, specimens can be sent directly to EMAGE for entry into the public database.

**Bioinformatics****Oral Presentation****Saturday November 5****5.15pm – 5.30pm****O-10****BEYOND THE DATA DELUGE: ONTOLOGIES FOR COMPLEX BIOLOGICAL SYSTEMS**

J A Blake, H J Drabkin, J A Kadin, L Ni, JE Richardson, M E Dolan, A Diehl, D P Hill  
The Jackson Laboratory, Bar Harbor, United States

The Gene Ontology is a structured vocabulary system that provides ontologies for functional annotations of all organisms. The GO system enables functional data analysis of very large data sets, e.g., the analysis of micro-array results. The GO annotations for mammalian genomes such as mouse, rat and human are comprehensive. The GO is an open source project; all ontologies and annotations and analysis tools are available at <http://www.geneontology.org/>

MGI is one of the founding groups of the GO and actively involved in the development and implementation of bio-ontologies. One of the major challenges at MGI is to create a human-digestible representation of the wealth of information in our database resource. To this end, we have development an ontology browser, a graphical representation of mouse/human/rat annotation sets for orthologous genes, and a textual representation of the structured annotations.

We are now exploring the power of disease-centric ontologies to support comparative analysis of mouse models and human disease presentations. This work draws on foundational ontologies such as the Gene Ontology (GO), the Anatomical Dictionary for Mouse, and the Mammalian Phenotype Ontology as well as other semantic standards incorporated into the Mouse Genome Informatics (MGI) system.

The Mouse Genome Informatics (MGI) Resource provides information about the genetics, and biology of the laboratory mouse. Community semantic standards facilitate the integration and recovery of mouse information, and the interconnection of mouse data with other biological information.

This work is supported by program project grant HG00330 and grant HG002273 from NHGRI, and grant HD33745 from NICHD.