THE NATIONAL CENTER FOR
BIOMEDICAL ONTOLOGY

Workshop on the Representation of Phenotypes

December 1st and 2nd, 2006 Clark Center, room S360, Stanford University

Thursday, November 30th, 2006: Dinner

7:30pm	Dinner at Saint Michael's Alley, 806 Emerson Street, (650.326.2430), http://www.stmichaelsalley.com/

Friday, December 1st, 2006: Driving Projects

8:40am	**Suzanna Lewis**: Background and motivation
9:00am	**Gully Burns**: Ontology engineering approaches based on semi-automated curation of the primary literature
9:20am	**Alexander Garcia**: The use of concept maps and automatic terminology extraction during development of a domain ontology: Lessons learnt.
9:40am	**Adrien Coulet**: Suggested ontology for pharmacogenomics (SO-Pharm): modular construction and preliminary testing
10:00am	**Tina Hernandez-Boussard**: Challenges for representing phenotype in pharmacogenomics
10:20am	Break (30 minutes)
10:50am	**James Reecy**: A project for the Creation of a Unified Trait Vocabulary for Farm Animals
11:10am	**Victoria Petri**: Rat genome database disease portals: a platform for genetic and genomic research
11:30am	**Carol A. Bastiani**: WormBase phenotype ontology development, curation, and data integration
11:50am	**Janan Eppig**: Ontologies and vocabularies supporting data integration: emphasis on mouse phenotypes and disease
12:10pm	Lunch
1:50pm	**Christopher D. Smith**: SIBO: A Social Insect Behavior Ontology for Ants and Bees
2:10pm	**Rex Chisholm**: Phenotype curation tool and ontologies at dictyBase
2:30pm	**Lindsay G. Cowell**: Immune systems
2:45pm	**Minna Lehvaslaiho**: *HIV*
3:00pm	**Christopher J. Mungall & Jim Hu**: E. coli and other bacteria
3:15pm	**Eurie Hong**: Yeast
3:30pm	**Tanya Berardini**: *Arabidopsis*
3:45pm	**Michael Ashburner**: *Drosophila*
4:00pm	**Maryann Martone**: BIRN
4:15pm	**Jim Balhoff**: NESCENT
5:30pm	Dinner at Fuki Sushi, 4119 El Camino Real, (650.494.9383), http://www.fukisushi.com/.

Saturday, December 2nd, 2006: Planning Ahead

9:00am	**Discussion:** commonalities, and consideration of discussion topics for the afternoon.
9:30am	**Christopher J. Mungall**: Representation of phenotype
10:30am	Break (30 minutes)
11:00am	**Mark Gibson**: Phenote hands-on
11:30pm	**Monte Westerfield**: Early results from PaTO and Phenote in use for Zebrafish
12:00pm	Lunch (1.5 hours)
1:30pm	**Issues of representation**: Extending Phenotype Entity-Quality representation
2:30pm	**Ontology Content**: Extending PaTO, and other key ontologies. Please bring use cases and we will interactively edit PaTO to suit requirements
3:30pm	Break (30 minutes)
4:00pm	**Annotation tools**: Extending Phenote capabilities, integration of other tools
5:00pm	**Review**: necessary follow-up, possible action items, and key decisions

Ontology engineering approaches based on semi-automated curation of the primary literature

Gully APC Burns, Tommy Ingulfsen, Donghui Feng, and Eduard Hovy

The process of knowledge curation from the primary literature is time-consuming, laborious, and specialized. Fortunately, the similarities of the curation process to the annotation of text with semantic labels presents an opportunity to employ cutting-edge natural language processing techniques to facilitate ontology construction. The result is that manual curation work can support both the development of an automated curation system and the semi-automated construction of a formal model of the domain. We present a general approach based on the use of active learning methods in conjunction with text-mining systems using the Conditional Random Fields model. Ultimately, we wish to construct annotation tools that fit seamlessly into scientists' everyday interaction with the primary literature. Our secondary, and complementary, focus is the creation of a domain ontology of the types of information identified for curation, which may encode formally the expert's knowledge and help pinpoint errors or vagueness in his or her understanding. We present preliminary data taken from information extraction experiments performed on the neuroanatomical connectivity literature. While this data is not normally considered a 'phenotype' within neuroanatomy, we argue that it (along with other non-genomic data) should be considered by the PATO community.

This work is funded by the Information Sciences Institute and the National Library of Medicine (LM-07061).

The use of concept maps and automatic terminology extraction during development of a domain ontology: Lessons learnt.

Alexander Garcia

Extracting terminology is not always an integral part within most methodologies for building ontologies. Moreover, the use of terms extracted from literature relevant to the domain of knowledge for which the ontology is being built has not been extensively studied within the context of knowledge elicitation. We present here some extensions to the methodology proposed by Garcia et al. (BMC Bioinformatics 7:267, 2006); two important advances on the initial proposed methodologies are the use of extracted terminology for framing the conceptual mapping building, and the use of narratives during the knowledge elicitation phases.

Suggested ontology for pharmacogenomics (SO-Pharm): modular construction and preliminary testing

Adrien Coulet, Malika Smaïl-Tabbone, Amedeo Napoli, and Marie-Dominique Devignes

Pharmacogenomics studies the involvement of interindividual variations of DNA sequence in different drug responses (especially adverse drug reactions). Knowledge Discovery in Databases (KDD) process is a means for discovering new pharmacogenomic knowledge in biological databases. However data complexity makes it necessary to guide

the KDD process by representation of domain knowledge. Three domains at least are in concern: genotype, drug and phenotype. The approach described here aims at reusing whenever possible existing domain knowledge in order to build a modular formal representation of domain knowledge in pharmacogenomics. The resulting ontology is called SO-Pharm for Suggested Ontology for Pharmacogenomics. Various situations encountered during the construction process are analyzed and discussed. A preliminary validation is provided by representing with SO-Pharm concepts some well-known examples of pharmacogenomic knowledge.

Challenges for representing phenotype in pharmacogenomics

Tina Hernandez-Boussard

The PharmGKB (http://www.pharmgkb.org/) is an online resource devoted to comprehensive cataloguing of genetic variations relevant to variation in drug response. We curate primary data (genotype, phenotype at molecular, cellular, clinical level) as well as knowledge (literature curation, pathways, human annotations of key genes). We provide search and visualization tools for this information, in order to catalyze research in pharmacogenomics. For both activities, we need to index the relevant phenotypes for the purposes of indexing, aggregation, search, and automatic summarization and data mining. We need a flexible method for annotating phenotypes that are described in the literature (by curators). We would prefer to adopt community-based standards that would allow PharmGKB to interoperate with other databases, both human and model organism.

A Project for the Creation of a Unified Trait Vocabulary for Farm Animals

LaRon Hughes, Jie Bao, Zhi-Liang Hu, Vasant Honovar, James Reecy

Ontologies help to identify and formally define the entities and relationships in specific domains of interest. Ontologies play a central role in annotating, integrating, analyzing, and interpreting biological data. Animal Trait Ontology (ATO) project being carried out under the auspices of the USDA-Animal Genome Research Program is aimed at developing a standardized trait ontology for farm animals and software tools to assist the research community in collaboratively creating, editing, maintaining, and using such ontology for annotating and querying data. Towards this end, we have developed a Collaborative Ontology Building (COB) editor to allow ATO curation by multiple, geographically distributed individuals and research groups. We have also developed database structures to maintain the ATO for cattle, pigs, chickens, and other livestock species. The ATO project extends our work on Animal QTLdb by adding new capabilities for standardization, annotation, retrieval, integration, and analysis of animal trait information, including in particular, QTL related traits.

Rat genome database disease portals: a platform for genetic and genomic research.

Victoria Petri, Mary Shimoyama, Andrew Patzer, Renee White, Jennifer Smith, Dorothy Reilly, Rajni Nigam, Weihong Jin, Jiali Chen, Simon Twigger, Anne Kwitek, Howard J. Jacob.

The Disease Portals at RGD provide a comprehensive research platform through the integration of heterogeneous datasets into the context of the genome using multiple ontologies and tools for data mining and visualization. The portals provide both the novice/experienced user with easy access to a comprehensive, integrated knowledgebase. Current/proposed components of the portals include: 1) comprehensive rat, human and mouse gene sets associated with diseases, related phenotypes, pathways and biological processes; 2) all rat QTLs related to a disease, associated mouse/human QTLs; 3) strains used as disease models; 4) phenotype data in a species-dependent manner; 5) references; 6) expression data; 7) genome-wide view of genes/QTLs via GViewer; 8) comparative maps of disease related regions, 10) customization of datasets/download options; 11) analysis/visualization of function and cellular localization makeup of gene sets. The portals are designed to highlight genetic/genomic data generated from rat research in diseases related to the cardiovascular, nervous, musculoskeletal, digestive, endocrine and immune systems as well as metabolic diseases, cancer. Disease data across the three species, along with species-dependent phenotypic data provide the user with a means to distinguish between subtle differences in disease manifestations. Such differences could help elucidate the links between events and conditions, the mechanisms that lead from the normal to the diseased phenotype.

WormBase phenotype ontology development, curation, and data integration

Carol A. Bastiani

WormBase integrates genomic, genetic, and a diverse array of biological data pertaining to C. elegans in a convenient format for browsing, directed searches, complex queries, and data downloads. There exists a wealth of phenotypic information in C. elegans derived from isolated mutants, RNAi experiments, and transgenic worms. Mass RNAi screens provide the bulk of RNAi-phenotype information, and WormBase serves as a central repository for this data. Prior to July, 2006, WormBase had been curating RNAi-phenotype information from mass screens and from individual RNAi experiments, extracted from the literature, using a phenotype vocabulary of about 120 terms. As of July, 2006, WormBase released a C. elegans phenotype ontology, currently comprised of over 1200 phenotype objects. WormBase has begun to curate phenotypic data from alleles and transgenes, as well as RNAi experiments, using the C. elegans phenotype ontology. In addition to associating phenotypes directly with relevant objects, WormBase is also associating phenotype objects using a "Not" qualifier, when it is communicated that a particular phenotype is not associated with an object. Currently, WormBase has made over 270,000 RNAi-phenotype and over 4,300 allele-phenotype associations, including "Not" associations. The WormBase phenotype ontology

represents a starting point from which we could eventually migrate to a PATO schema for describing phenotypes.

Ontologies and vocabularies supporting data integration: emphasis on mouse phenotypes and disease

Janan T. Eppig, Susan M. Bello, Judith A. Blake, James A. Kadin, Martin Ringwald, Kim Forthofer, Richard M. Baldarelli, Cynthia L. Smith.

The mouse is an exceptional model system for connecting knowledge from sequence-to-phenotype-to-disease. The Mouse Genome Informatics Database (MGI, http://www.informatics.jax.org) supports biological knowledge building for mouse by integrating genetic, genomic, and biological data and facilitating data mining and complex querying. Full access to integrated data is enabled by extensive use of structured vocabularies and ontologies including the Gene Ontology (GO), mouse Embryonic and Adult Anatomical Dictionaries (EMAP and MA), Mammalian Phenotype (MP) Ontology, and Online Mendelian Inheritance in Man (OMIM) disease and syndrome terms. In addition, MGI is the authoritative source for nomenclature for mouse genes, alleles, and strains. Many smaller vocabularies, such as mutation class, sequence type, genetic marker type, expression assay type, etc., also are key to MGI data integration. Phenotypic descriptions in MGI rely on the MP Ontology and definition of specific genotypes and strain backgrounds. The MP Ontology has been adopted successfully to describe mouse (MGI), rat (RGD), human (NBCI), and animal (OMIA) phenotypes. As of July 2006, MGI included >16,000 alleles representing phenotypic mutations in >6,600 genes. Over 65,600 phenotype annotations in MGI have been made using MP Ontology terms. The MP Ontology itself has, thus far, grown to >4,400 defined terms. Over 1,700 mouse models are associated with OMIM disease terms. Supported by NIH grant HG00330.

SIBO: A Social Insect Behavior Ontology for Ants and Bees

Christopher D. Smith. Dept. of Biology & Center for Computing for Life Sciences. San Francisco State University, San Francisco, CA 94132.

Despite mounting evidence for the role of genes in influencing complex behaviors the molecular pathways and mechanisms remain poorly understood. Social insects are an excellent evolving model system for investigating the roles between genes and behavior because genetically identical individuals can exhibit dramatically different behaviors based on their caste or age. For example, *young* ants and honeybee workers typically tend the nest and young, while *older* members of the nest take on riskier activities like defense and foraging. Social insects have a rich natural history with many well-described behaviors in the literature. Furthermore, recent microarray studies have shown that workers and queens differentially express a number of caste-specific genes and studies in progress promise to find even more associations between specific genes and behaviors like foraging, aggression, and nest care.

Linking genes to behaviors will require extensive integration of behavioral observations and functional genomic data. We propose to catalog and formalize this extensive

behavior repertoire in the Social Insect Behavior Ontology (SIBO). Since behavior and communication are inextricably linked, we will also adapt existing anatomy and chemical ontologies and curate the specific relationships between communication molecules, the anatomies that produce them, and the behaviors associated with them. SIBO will require extensive literature curation and will initially focus on curating honeybee and ant genes linked to behaviors from high-throughput experiments and identifying orthologs of *D. melanogaster* behavior genes described in the Gene Ontology (GO).

Phenotype curation tool and ontologies at dictyBase

Rex L. Chisholm, Warren A. Kibbe, Pascale Gaudet, Karen E. Pilcher, Petra Fey, Sohel Merchant, Eric Just

Phenotype is an important functional annotation available at dictybBase. Although in the past we have associated phenotypes to genes, it is more accurate to represent them as characteristics of a strain. Therefore we have transitioned towards associating phenotypes to strains and using PATO together with a species specific ontology to describe phenotypes. This involved the development of a species-specific phenotypic ontology, and software tools required to support the use of these ontologies for the phenotypic annotation of strains. Our interim phenotype ontology contains 415 terms and is based on the Entity-Quality model, where each term consists of an entity, which is either a GO term or a Dictyostelium anatomy term, and a quality. Examples of Dicty Phenotype Ontology terms include: aberrant fruiting body morphology, decreased exocytosis, and aberrant cytoskeleton organization. As PATO becomes more stable, we will compose these terms using PATO qualities. We have developed a web-based software tool to curate phenotypes that runs on top of the Chado schema from GMOD. This tool has been developed in collaboration with NCBO. It is very similar in functionality to Phenote and provides the curators the ability to capture data using the Entity-Quality model. In addition to the phenotype described by the entity and quality, the tool allows the curators to capture assay, environment, reference, and genetic context.

Supported by NIH grants GM064426 and HG00022.