

## ISMB 2005 Tutorial

### Semantic Aggregation, Integration, and Inference of Pathway Data

or

*"Pedantic Aggravation, Irritation, and Interference of Pathologic Detritus."*

**Instructors: Joanne Luciano and Jeremy Zucker**

#### **Tutorial Level: Intermediate**

#### **Expected Goals and Objectives**

The objective of this tutorial is provide the student with the knowledge and tools necessary to perform semantic aggregation, integration and inference on biological pathway data.

#### **Motivation**

The Pathway Resource List at <http://cbio.mskcc.org/prl> contains over 170 biological pathway databases, and the list continues to grow. However, in order to consolidate the knowledge required for many research projects, one must extract the relevant pathway data from each database, transform it into a standard data representation, and load it into an integrated repository.

The motivation for this tutorial is to enable more productive and efficient pathway-based research.

Knowledge aggregation, integration, and inference issues such as data cleaning, semantic mapping, and fundamentals of the extract, transform and load (ETL) methodology will be illustrated through real world examples of biological pathway data.

#### **Intended Audience**

The intended audience consists of bioinformaticians, computational biologists and database developers. Participants should be familiar with

- intermediate-level programming concepts (APIs, UML, XML parsing, object models)
- intermediate-level database concepts (SQL, data modeling, Entity-Relation diagrams, etc.)

Topics covered include biological knowledge representation issues, data cleaning, and fundamentals of the extract, transform and load (ETL) methodology, illustrated through real world examples of semantic aggregation, integration and inference.

#### **Format of the tutorial:**

Introduction (45 minutes)

Workshop (2 hrs 15 minutes total, 45 minutes each)

    Subdivide into groups of triads and dyads

    3 Case Studies (lecture) with corresponding in-class exercises (small groups)

Lessons Learned (30 minutes)

Lessons Not Yet Learned (take home exercise)

## Outline of the Tutorial

**"What is a Pathway? It depends on who you ask." - Joanne Luciano**

**"A good representation is the key to good problem solving" –Patrick Winston**

### Introduction: Biological Knowledge Representation Background

**Objective** Each pathway modality has its own specific representation issues which must be understood before attempting to integrate across modalities.

**Metabolic pathways** describe the network of enzyme-catalyzed reactions that release energy by breaking down nutrients (catabolism) and building up the essential compounds necessary for growth (anabolism). Experimentally-determined metabolic pathways have been established for a few model organisms, but most metabolic pathways databases contain pathway data that has been computationally inferred from the genome annotations. Because most genome annotations are incomplete, metabolic pathway databases contain pathway holes which can only be addressed by experiment or computational inference.

A good test of a reconstructed metabolic network is to ask if it can produce the set of essential compounds necessary for growth, given a known minimal nutrient set. To solve this problem, metabolism can be represented as a bipartite directed graph, where one set of nodes represents metabolites, the other set represents biochemical reactions and labeled edges are used to indicate relationships between nodes (reaction X *produces* metabolite Y, or metabolite Y *is-consumed-by* reaction X). Now, the biological question can be answered by computing the *transitive closure* of the reactions that can be fired to see if the essential compounds are produced given the minimal nutrient set.

**Gene regulatory networks** describe the network of transcription factors that bind regulatory regions of specific genes, and activate or repress their transcription. Gene regulatory networks, or transcription networks, have been found to contain recurring biochemical wiring patterns, termed network motifs, which carry out key functions. When nature chooses one kind of circuit over another, it is often because that circuit displays properties that are advantageous. For example, the feed-forward loop (FFL) is a three-gene pattern composed of two input transcription factors, one that regulates the other, both jointly regulating a target gene. In the case of the feed-forward loop, this motif enables the organism to have finer control over the dynamics of the target gene. Therefore, finding statistical irregularities in the predominance of one kind of network motif over another indicates a possibly important function. How does one find the most significant recurring network motif in a given transcriptional network? To answer this question, transcription networks can be described as directed graphs, in which the nodes are genes, and edges represent transcription interactions, where a transcription factor encoded by one gene modulates the transcription rate of the second gene. By performing multiple network motif search queries over large numbers of randomized graphs and real network, it is possible to calculate statistical distributions that can then be compared against the motif distribution of one target biological network. Overrepresented motifs indicate a functional advantage for the organism, and give clues for constructing the dynamics of a genetic regulatory model of the organism (Mangan 2003).

*Clique, paraclique* and other forms of *densely-connected subgraphs* are also useful in isolating sets of putatively co-regulated genes, which can be an important first step in gene regulatory network elucidation.

**Signaling pathways** describe biochemical reactions for information transmission and processing. Unlike metabolic pathways that catalyze small molecule reactions, signaling pathways involve the post-translational modification of proteins, leading to the downstream activation of transcription factors. [Put in a sentence here that explains that there are many different post translational modifications, what the most popular ones are (so that when the students hear them they'll know what's being talked about) and

then a reference to Ken Fukuda's ontology for a complete list.

To infer these signal transduction networks from protein-interaction data, one can select all linear pathways starting from any membrane protein and ending at any DNA-binding protein, use microarray data to rank each pathway according to the co-expression of its pathway constituents, and combine pathways with similar start and end points to produce the final signal transduction pathway (Steffen 2002).

Integration of signaling pathways poses a greater challenge than with metabolic pathways because of the diversity of representation schemes for signaling. Some signaling databases, such as PATIKA and INHO, use compound graphs to represent signaling pathways, while other object-oriented databases use inheritance to establish relationships between post-translational modifications of proteins.

Because many kinetic parameters are unknown for many signal transduction networks, and many of the protein species are available in low copy numbers, it is difficult to model the dynamics with ordinary differential equations. Stochastic methods have been combined with knowledge-based formalisms to model and simulate the dynamics. (PathwayLogic), (BioSigNet), (BioNetGen), (BioSPI), (Valis)

**Protein interaction networks** describe protein-protein, protein-DNA, protein-RNA or protein-ligand bindings that have been experimentally observed. One problem that arises in protein-protein interactions is that the network data is notoriously unreliable. To extract true interactions from the false positive and false negative rates, one can represent the data as undirected graphs to record pairs of proteins that are experimentally observed to interact with each other. Then, queries consisting of Boolean graph operations, i.e., graph intersection, majority and at-least-k-of-n over multiple graphs can be used to refine the data. Once the data has been cleaned, one can discover uncharacterized functional modules by looking for conserved protein interaction pathways using pathway alignment (Kelley, et al. 2004), or by overlaying expression data (Vidal), infer signal transduction pathways by looking for shortest paths between receptors and transcription factors (Steffen), identify members of a enzyme complex by integrating with metabolic pathway annotations, etc.

**Taxonomies of proteins, chemical compounds, and organisms** classify their instances according to functional, structural and/or phylogenetic properties. These classification systems are usually represented as directed acyclic graphs (partial orders or lattices) in contrast to the network datasets described above that often contain cycles. Such taxonomies are commonly used in querying network databases, e.g., to restrict the allowable label to be a member of class of enzymes (such as kinase enzymes, rather than a specific label such as ERK1. Because taxonomies never have cycles, they are amenable to more efficient algorithms.

**Chemical structure graphs** use nodes to represent atoms undirected edges to represent chemical bonds. Chemical structure graphs are widely used in chemical information retrieval systems such as Chemical Abstracts Service. Common queries include chemical substructure matching. Chemical graphs can be integrated with metabolic pathway databases under a stratified nested graph model to permit chemical substructure matching (for chemical entities) as part of larger queries against biological networks. BioCyc, Klotho and LIGAND support chemical structure graphs.

# Case Study I

## Semantic aggregation of Pathway Data

### Motivation

The marine cyanobacteria, *Prochlorococcus marinus*, are the smallest and most abundant photosynthetic organisms. These organisms are of great ecological importance; along with other marine cyanobacteria from the *Synechococcus* genus, these cells are among the primary producers of the phytoplankton, which are responsible for half of the photosynthesis on the planet. These bacteria are therefore prominent actors in global oceanic function, in the carbon cycle, and consequently in the evolution of climate. [http://www.genoscope.cns.fr/externe/English/Projets/Projet\\_DR/organisme\\_DR.html](http://www.genoscope.cns.fr/externe/English/Projets/Projet_DR/organisme_DR.html)

One hypothesis, known as the "Iron Hypothesis," suggests these organisms may hold a possible solution to global warming <http://www.agu.org/revgeophys/chisho00/chisho00.html>. It is clear these are an important organism for us to understand.

### Goals

**Integrated analysis** In order to be able to understand microarray data and other kinds of large-scale measurements, it is essential to have the information about each gene and gene product available at the biologist's fingertips. Semantic aggregation challenge: Aggregate RNA and protein expression data to constrain the reaction fluxes of the metabolic model.

Constrain metabolic flux model with experimental measurements:

- RNA expression
- Protein expression
- Metabolite concentrations
- Flux measurements

### Data Sources

Public:

- BioCyc (metabolism)
- TransportDB (transport proteins)

Local:

- RNA expression (microarrays)
- protein expression (mass spec)

### Workshop exercise:

Learn how to solve a data aggregation problem using semantic web technologies.

LSID's, OWL/RDF, UnificationXref, RelationshipXref

Examples: recreate Siderean demo, BioDash demo, *P. marinus* demo

**Discussion** of Schema-level errors, instance-level errors, testing and validation solutions.

## Case Study II

### Semantic Integration of the E. coli metabolic flux model

"1+1 is equal to, but not identical to 2"—Author not yet identified.

"Like and equal are not the same thing at all"—A Wrinkle in Time

#### Motivation

Given two metabolic databases, A and B of the same organism, which metabolite in A equals which metabolite in B? Solution: semantic mapping and data merging.

#### Goals

- Best of both worlds
- Biological Objective: From nutrients create all essential compounds required for growth
- True test of metabolic databases: Is the data good enough to predict growth rate under different nutrient conditions and effect of gene knockouts?

#### Data Sources

##### **EcoCyc - Literature-derived metabolic pathway/genome database of E. coli**

- good schema
- Flux balance model doesn't work

##### **JR904 – Literature derived flux balance model of E. Coli**

- implicit schema
- good metabolic flux balance model of E. coli
- literature curated biochemical reactions of E. coli
- 904 enzymatic reactions
- gene, enzyme-reaction associations

#### Software Tools

Mapping and merging tools with the semantic web.

"How to make your data integration project useful to everyone else"

#### Workshop exercise:

- mapping from BioCyc ontology to BioPAX ontology
- mapping of implicit JR904 schema to BioPAX ontology
- aggregation of JR904-specific concepts (flux limits) with BioPAX concepts

## Mapping Issues

- EcoCyc <-> JR904 Gene names
- EcoCyc <-> JR904 Enzyme names
- EcoCyc <-> JR904 Reaction names
- EcoCyc <-> JR904 Reversibility/flux limits
- EcoCyc <-> JR904 Gene->protein associations
- EcoCyc <-> JR904 protein->enzyme complex associations
- EcoCyc <-> JR904 enzyme->reaction associations

## Discussion

- Definitions of Error, inconsistency, discrepancy,
- Bugs you can find by comparing data with schema constraints (mass balance, Central Dogma) (logical inconsistency)
- Bugs you can only find by comparing one data set with another data set (primary source) (data inconsistency)
- Bugs you can only find by comparing data with experiment. (errors in underlying assumptions) (publish a paper, we have a new biological concept!)

## Case study III:

### Semantic Inference of a Pathway/Genome Database using PathoLogic

#### Motivation

Science is about discovering new knowledge by rigorously examining the underlying assumptions in our concepts of reality. Each pathway data model contains a formal representation of those concepts, thereby making those assumptions explicit.

**Constraint-based modeling** Flux Balance Analysis (FBA) is an approach to modeling the metabolic network of an organism by using the stoichiometric coefficients and thermodynamic reversibility to constrain the metabolic network. Kinetic information is not required. Optimization of an objective function, such as growth rate, is used to obtain a metabolic flux distribution that satisfies the constraints. FBA has been shown to provide meaningful predictions in *Escherichia coli* and *Saccharomyces cerevisiae*. Semantic inference challenge: Given an annotated genome, is it possible to reconstruct the entire metabolic network?

#### Goals

Reconstruction of an organism's metabolic network in order to perform in silico experiments of the networks behavior that can be tested and verified experimentally.

#### Sub goals

1. Infer genes from sequence data and sequence homology.
2. Infer enzymatic reactions from gene annotation Enzyme Commission (EC) numbers
3. Infer metabolic reaction network from enzymatic reactions and metabolites.
4. Infer pathway holes using network debugging algorithms
5. Propose candidate enzymes using pathway hole-filling algorithms.
6. Add experimentally verified candidates to the annotated genome.
7. Lather, rinse, repeat.

#### Workshop exercise:

Manually infer a metabolic pathway from a partially-annotated mini-genome

#### Discussion of schema and instance-level errors:

- Relationships between genes proteins complex reactions and pathways
- Mapping genbank tag to BioCyc ontology
- Two reactions same EC number
- naming problems (synonyms)
- errors in names
- errors in the values associated with those names
- unbalanced reactions

## Lessons Learned

*"When it comes to data cleaning, there is no such thing as a free lunch."*

*-- Sir Tim Berners-Lee*

***"Above all, one must have a feeling for the organism" –Barbara McClintock***

Integrative systems biology combines efforts from biologists, mathematicians, computer scientists, and engineers to analyze the network of interacting parts that make up biological systems. As in any interdisciplinary subject, a common language must arise that allows its practitioners to communicate with one another. One anticipated result of this communication is powerful new biological insights that could not otherwise have been obtained alone from any individual discipline.

## References

1. Mangan S, Alon U. "Structure and function of the feed-forward loop network motif." *Proc Natl Acad Sci U S A*. 2003 Oct 14;100(21):11980-5. Epub 2003 Oct 6.
2. Steffen M, Petti A, Aach J, D'haeseleer P and Church G. "Automated modelling of signal transduction networks" *Bioinformatics* 2002, 3:34
3. Segre D, Zucker J, Katz J, Lin X, Dhaeseleer P, Rindone W, Karchenko P, Nguyen D, Wright M, and Church GM (2003) "From annotated genomes to metabolic flux models and kinetic parameter fitting." *Omic*s 7:301-16.
4. Luciano, JS "PAX of Mind for Pathways Researchers" *Drug Discovery Today* (tentative July 2005).

## Resources:

1. <http://cbio.mskcc.org/prl>
2. [www.biocyc.org](http://www.biocyc.org)
3. [www.biopax.org](http://www.biopax.org)
4. **Iron Hypothesis:** <http://www.agu.org/revgeophys/chisho00/chisho00.html>
5. Cary, M. P. *et al.* (2005) *Pathway information for systems biology*. 579:1815-1820, 2005.
6. Karp, P. D. *et al.* (2002) *The MetaCyc Database*. 30: 59-61, 2002.
7. Bader, G. D. *et al.* (2003) BIND: the Biomolecular Interaction Network Database. *Nucleic Acids Res.*, 31: 248-250, 2003
8. Murray-Rust, P. and Rzepa, H. S. (2003) Chemical markup, XML, and the World Wide Web. 4. CML schema. *J. Chem. Inf. Comput. Sci.*, 43: 757-772, 2003
9. Weininger, D. (1988) SMILES, a Chemical Language and Information System. 28: 31-36, 1988.

10. Hermjakob, H., *et al.* (2004) The HUPO PSI's molecular interaction format--a community standard for the representation of protein interaction data. *Nat. Biotechnol.*, 22: 177-183, 2004.
11. Hucka, M., *et al.* (2003) The systems biology markup language (SBML): a medium for representation and exchange of biochemical network models. *Bioinformatics*, 19: 524-531, 2003.
12. Spyns, P., Meersman, R., and Jarrar, M. (2002) Data Modelling versus Ontology Engineering, *SIGMOD Special Section on Semantic Web and Data Management* 31: Number 4, December 2002
13. Lemer, C., *et al.* (2004) The aMAZE LightBench: a web interface to a relational database of cellular processes. *Nucleic Acids Res.*, 32 Database issue: D443-448, 2004.
14. Karp, P. D., *et al.* (2002) The EcoCyc Database. *Nucleic Acids Res.*, 30: 56-58, 2002.
15. Overbeek, R., *et al.* (2000) WIT: integrated system for high-throughput genome sequence analysis and metabolic reconstruction. *Nucleic Acids Res.* 28: 123-125, 2000.
16. Kanehisa, M., *et al.* (2004) The KEGG resource for deciphering the genome. *Nucleic Acids Res.*, 32 Database issue: D277-280, 2004.
17. Joshi-Tope, G., *et al.* (2005) Reactome: a knowledgebase of biological pathways. *Nucleic Acids Res.*, 33 Database Issue: D428-432, 2005.
18. The Gene Ontology Consortium (2000) Gene ontology: tool for the unification of biology. *Nature Genet.* 25: 25-29, 2000.
19. Rumble, J. Jr., Lee, A. Y., Blakeslee, D., and Young, S. (2001) Reliable solubility data in the age of computerized chemistry. Why, how, and when? *Pure Appl. Chem.* 73: 825-829, 2001.
20. Noy, N., and McGuinness, D. *Ontology Development 101: A Guide to Creating Your First Ontology*. Stanford University, Stanford, CA, 94305.
21. Horridge, M. *et al.* (2004) *A Practical Guide to Building OWL Ontologies Using The Protege-OWL Plugin and CO-ODE Tools*. Edition 1.0 The University Of Manchester. August 27, 2004.

## Instructors

### Joanne S. Luciano, PhD



Joanne Luciano is an active leader in several community-based initiatives, including BioPAX, the BioPathways Consortium, and the emerging Semantic Web for Life Sciences. She co-developed the BioPAX ontology that is on its way to becoming the standard for biological pathway knowledge representation. She is an authority in pathway databases and modelling languages. Joanne has been an active member of the computational and systems biology community since 1996, presenting at many international conferences. Joanne holds a joint appointment at Harvard Medical School and Massachusetts General Hospital. She also is President and Founder of Predictive Medicine, Inc and holds two US patents for her analysis methods.

### Jeremy Zucker



<http://osgoodphoto.com/headshot/zucker.html> Jeremy Zucker holds degrees in Computer Science and Applied Mathematics from the University of Colorado. He is an expert in semantic aggregation, integration and inference. Currently a bioinformatics specialist at the Dana-Farber Cancer Institute and a computational biologist for the Church lab at Harvard Medical School, Jeremy is a lead developer for projects such as DARPA's BioSPICE, the DOE Genomes to Life, and BioPAX.