

ISMB 2005 Tutorial Proposal

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 150 biological pathway databases and is growing. However, to consolidate all the knowledge for a particular organism, one must extract the pathways from each database, transform those pathways into a standard data representation, and load the data into an integrated repository. The motivation for this tutorial is to enable more productive and efficient research by reducing the amount of work that a bioinformatician must perform in order to accomplish pathway integration.

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.

Outline of the tutorial:

“A good representation is the key to good problem solving” –Patrick Winston

Use cases for integrated pathway resources:

Objective: To provide the student with motivation for tackling data integration.

- Quantitative and Qualitative modeling (metabolic flux analysis)
- Comparative analysis (across multiple organisms)
- Integrated analysis (across different pathway modalities)

Network Inference (nutrient analysis)
Pathway hole-filling (finding orphan enzymes)
Network reconstruction (from annotated genomes to metabolic flux models)
Data mining (graph queries, relational queries, logic programming)

“What is a Pathway? It depends on who you ask.” - Joanne Luciano

Biological Knowledge Representation Issues (with examples)

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

Metabolism
Signal Transduction
Protein-protein interaction
Gene regulation
Genetic interaction
Protein-compound

“The great thing about standards is that there are so many from which to choose”

Objective: To demonstrate how to write interoperable software to navigate the alphabet soup of standards.

Metabolic Pathways: BioPAX Level 1
Models: SBML, CellML
Chemical: CML, inchi, SMILES,
Molecular interaction: PSI, BioPAX level 2
Sequence: SO
Gene: GO

Extraction, Transform, Load (ETL) methodology:

Data extraction methods:

SQL, flatfile parsing, database API's

Transformation into abstract data model:

UML, OWL/RDF, ER diagrams, XML, Object model.

Loading into data repositories:

Federated verses warehouse verses semantic web.

Reality check? Data cleaning!

Single-source problems

Schema level (integrity constraints)
Instance level (duplication, errors, inconsistencies)

Multi-source problems

Schema level (translation, integration)
Instance level (merge/purge, object identity problem)

“Above all, one must have a feeling for the organism” –Barbara McClintock

Case studies:

Single-source inference case study: Metabolic reconstruction

Pathologic: Annotated genome to pathway/genome database
BioCyc2SBML: Pathway/genome database to metabolic flux model

Dual-source integration case study: EcoCyc/JR904 integration

Integration Goal: resolve pathways, reactions, metabolites, genes, and enzymes
EcoCyc: Literature-derived metabolic pathway/genome database of E. coli
JR904: Literature-derived metabolic flux model of E. coli

Multi-source aggregation case study: BioSPICE/GtL use case: Debugging the Bug

Data source modalities: Genome, RNA levels, Protein levels, Metabolite profiles
Data repository: BioCyc/BioLingua/Biowarehouse/BioPAX/SBML

“Standard is Better than best” –Gerald J Sussman

Relevant qualifications and teaching experience of 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's teaching experience ranges over 20 years, and includes such courses as artificial intelligence, data communications and networks, and data structures. Joanne has been an active member of the computational and systems biology community since 1996, presenting at many international conferences.

Jeremy Zucker

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. His teaching experience includes a graduate course on systems biology at the Harvard Department of Molecular Cell Biology, and is a contributing author of the soon-to-be-released textbook: Introduction to Systems Biology (Science of Knowledge Press).