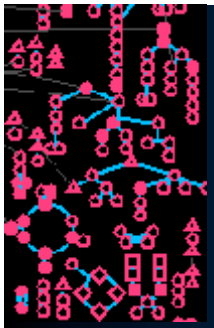


Databases and Algorithms for Pathway Bioinformatics

Peter D. Karp, Ph.D.

Bioinformatics Research Group

SRI International

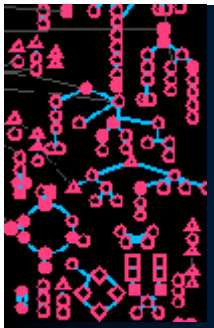
pkarp@ai.sri.com

BioCyc.org

EcoCyc.org

MetaCyc.org

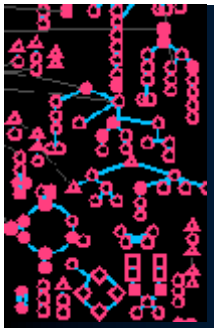
HumanCyc.org



MOD Home Pages

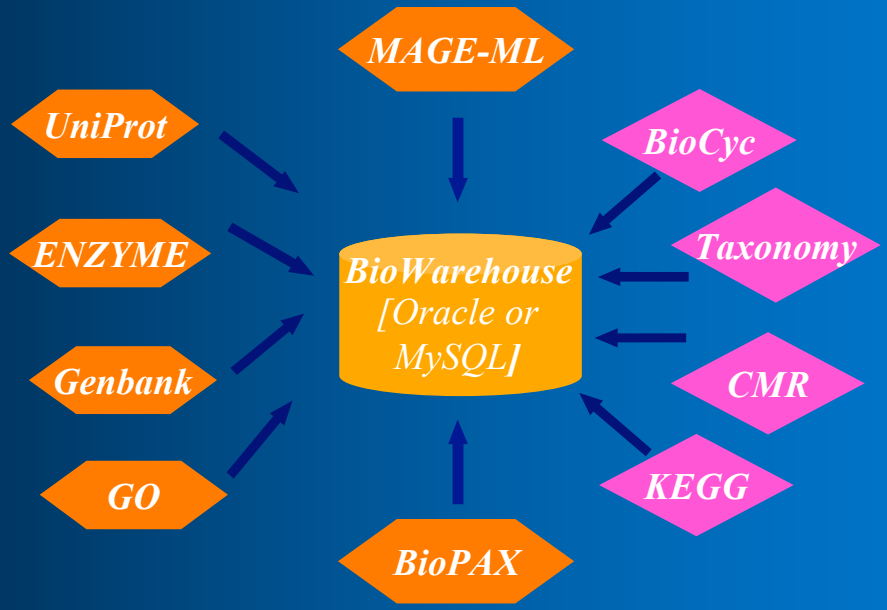
- **Learn or standardize?**

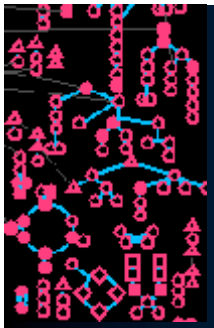
- **Top / Left**
- **Cascading menus or not**
- **Must-haves on home page:**
  - Citing MOD
  - Software/data download
  - Contact us
  - News
  - Publications
  - Statistics
  - Update history
  - Credits



BioWarehouse

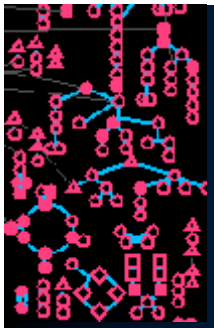
Peter D. Karp, Tom J. Lee,
Valerie Wagner, Yannick Pouliot





Motivations

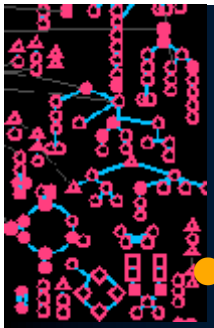
- Hundreds of bioinformatics DBs exist
- Important problems involve queries across multiple DBs



Why is the Multidatabase Approach Alone Not Sufficient?

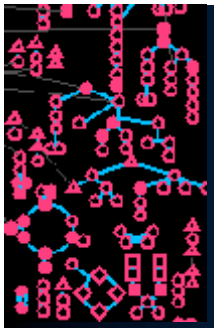
- Multidatabase query approaches assume databases are in a queryable DBMS
- Most sites that do operate DBMSs do not allow remote query access because of security and loading concerns
- Users want to control data stability
- Users want to control speed of their hardware
- Internet bandwidth limits query throughput
- Users need to capture, integrate and publish locally produced data of different types

- Multidatabase and Warehouse approaches complementary



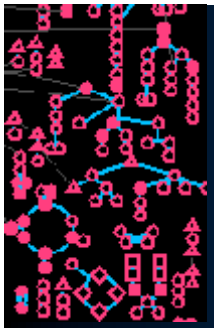
Technical Approach

- **Multi-platform support: Oracle (10g) and MySQL**
- **Schema support for multitude of bioinformatics datatypes**
- **Create loaders for public bioinformatics DBs**
  - Parse file format of the source DB
  - Semantic transformations
  - Insert DB contents into warehouse tables
- **Provide Warehouse query access mechanisms**
  - SQL queries via ODBC, JDBC, OAA
- **Operate public BioWarehouse server: publichouse**



BioWarehouse Schema

- **Manages many bioinformatics datatypes simultaneously**
  - Pathways, Reactions, Chemicals
  - Proteins, Genes, Replicons
  - Sequences, Sequence Features
  - Organisms, Taxonomic relationships
  - Computations (sequence matches)
  - Citations, Controlled vocabularies
  - Links to external databases
- **Each type of warehouse object implemented through one or more relational tables (currently 43)**



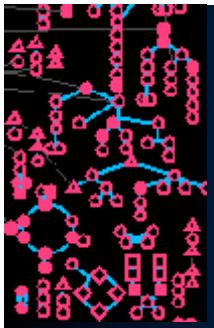
Warehouse Schema

- **Different databases storing the same biological datatypes are coerced into same warehouse tables**
- **Design of most datatypes inspired by multiple databases**
- **Representational tricks to decrease schema bloat**
  - Single space of primary keys
  - Single set of satellite tables such as for synonyms, citations, comments, etc.

BioWarehouse Loaders

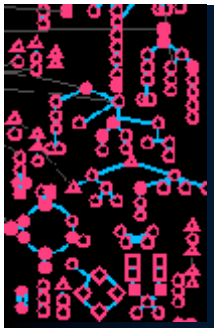
SRI International
Bioinformatics

Database	Loader Language	Input Format	Comments
Any BioPAX	Java	BioPAX	Protein interaction data only
BioCyc	C	BioCyc attribute-value	Pathway/Genome Databases
CMR	C	CMR column-delimited	Comprehensive Microbial Resource: 150+ microbial genomes
ENZYME	Java	ENZYME attribute-value	Enzyme Commission set of reactions
Genbank	Java	XML derived from ASN.1	Bacterial subset of Genbank
Gene Ontology	Java	OBO XML	Hierarchical controlled vocabulary
KEGG	C	KEGG format	Metabolic pathway data
Any MAGE-ML	Java	MAGE-ML format	Microarray gene expression data
NCBI Taxonomy	C	Taxonomy format	Organism taxonomy
UniProt	Java	UniProt XML	SWISS-PROT and TrEMBL

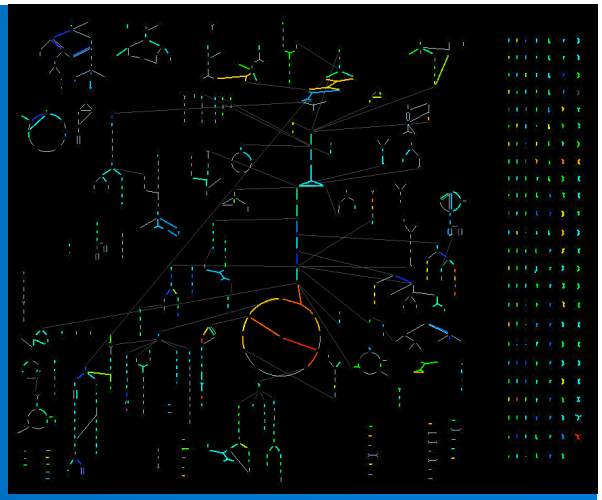


Uses of BioWarehouse

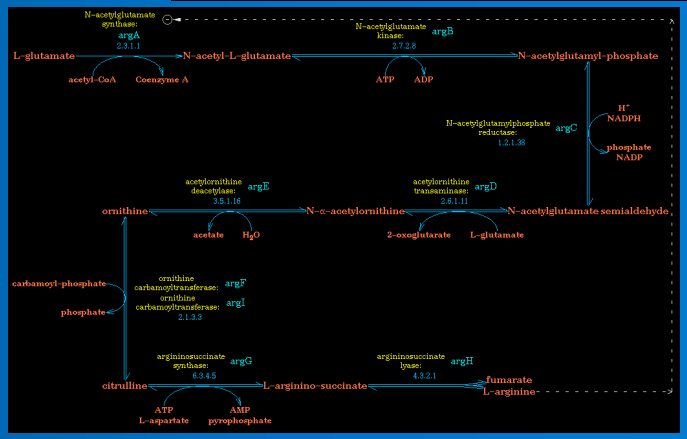
- **SRI Bioinformatics Research Group**
  - Extract genome data from CMR for generation of BioCyc Tier 3
  - Use CMR for genome-context extensions to pathway hole filler
  - Enzyme genomics research

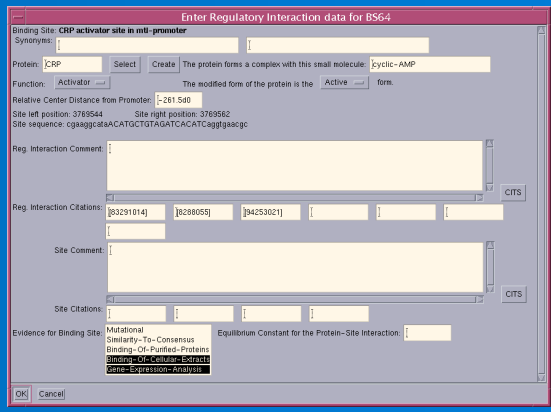


Pathway Tools Software



- **PathoLogic**
  - Predicts operons, metabolic network, pathway hole fillers, from genome
  - Computational creation of new model organism databases
  - Can work as complete MOD system, or pathway module only

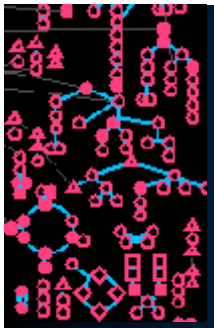




- **Pathway/Genome Editors**
  - Distributed curation of PGDBs
  - Distributed object database system, interactive editing tools

- **Pathway/Genome Navigator**
  - WWW publishing of PGDBs
  - Querying, visualization of pathways, chromosomes, operons
  - Pathway visualization of omics data
  - Global comparisons of metabolic networks

Brg.ai.sri.com/ptools/

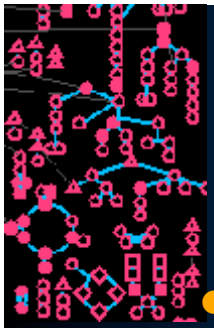


Pathway Tools Update

SRI International
Bioinformatics

- Version 10.5
- Ontology upgrade for signaling interactions
- Consistency checker
- Generate metabolic map poster
- Zooming in Overview
- Sequence retrieval tool
- BioPAX export, pathway page
- New reaction editor
- Spelling checker

- Version 11.0
- Regulatory network viewer



Acknowledgements

- **SRI**
  - **Suzanne Paley, Michelle Green, Ron Caspi, Ingrid Keseler, Mario Latendresse, Carol Fulcher, Markus Krummenacker, Alex Shearer**

- **Funding sources:**
  - NIH National Institute of General Medical Sciences
  - NIH National Center for Research Resources
  - NIH National Human Genome Research Institute

BioCyc.org

Learn more from BioCyc webinars: biocyc.org/webinar.shtml