Biopathways and Protein Interaction Databases

Frank Olken

Lawrence Berkeley National Laboratory

PGA Course at LBNL

Friday, Nov. 8, 2002

Protein Interaction Databases

- Protein interactions from experiments
- Used to help construction biopathways DB
- Yeast 2 Hybrid Experiments
  - pairwise interactions
- Mass spectrometry experiments
  - indentifies protein complexes
  - NOT pairwise interactions
- PIN = Protein Interaction Network

Biopathways Databases

- Metabolic pathways
- Signaling pathways
- Gene regulatory networks
- Inferred from:
  - protein interaction networks
  - micro-array data
  - other experiments
- BP = Biopathways

Copyright 2002 UC Regents

Outline of Talk

- Examples of BP and PIN Databases
- DB contents
- DB Uses
- Graph data model
- Graph queries
- Scale free networks

Copyright 2002 UC Regents

Database Contents

- **Networks of chemical reactions**
- **Metabolic: bulk reactions, ODE's**
- **Signaling: rarer reactants, stochastic Petri nets**
- **Gene regulatory networks**
  - **gene expression**
- **Protein Interaction networks**
  - **pairwise protein interactions**
  - **data is very noisy**

Copyright 2002 UC Regents

Directed vs. Undirected Graphs

- Biopathways = networks of chemical reactions
  - **also gene expression …**
  - **directed graphs**
  - **nodes = reactions or chemical entities**
  - **edges = relationships (inputs, outputs, catalyst …)**
- Protein interaction networks
  - **undirected graphs**
  - **nodes = proteins**
  - **undirected edges connect interacting proteins**

Copyright 2002 UC Regents

Pathway vs. Path

- Pathway:
  - typically a small connected subgraph of a larger graph (e.g., the entire metabolism of a microbe)
  - biological term
- Path:
  - a connected linear graph, i.e., no branches or cycles
  - A---->B---->C----->D
  - a term from graph theory

Uses of Biopathways DB

- Formal encoding of biological knowledge
- Drive simulations (add math models)
- Access to literature (links from reactions)
- Access to data (microarray data)
- Assist gene annotation
  - coregulation of genes (e.g., from microarray data) suggests participation in same pathways

Comparative Biology on Biopathways

- Comparative analyses of DNA/RNA/protein sequences has proven very useful
- Biopathways datasets for many organisms are becoming available
- We can now begin to do comparative analyses of biopathways
- Formal encoding of biopathways needed to permit automated comparisons

Comparative Biology of Sequences

- Sequences are abundant - gigabytes in Genbank
- Sequences are cheap to get now - pennies per base
- Sequences have been collected in databases
- Lots of analysis software available

Comparative Biology of Protein Structures

- Protein structures are expensive to determine - thousands of dollars each
- Few are known - 20K in PDB, 10K distinct
- Some software to compare
- Expensive computations
- Collected in PDB database

Comparative Biology of Biopathways

- Small datasets
  - presently dozens of organisms (mostly microbes)
  - soon hundreds of organisms
  - complete ???
- Expensive to generate
  - often requires micro-array experiments, wet chemistry, gene knockouts, …
- Multiple databases
- Limited analysis software

Requirements for pathway comparative biology

- Data (collections of biopathways)
- Data encoding (graphs)
- Algorithms for graph comparison
- Encodings for graph patterns
- Algorithms for graph pattern matching
- Tools for visualization of graph matchings

Copyright 2002 UC Regents

Comparative Biology Tools Sequences vs. Pathways

- Global alignment
- Local alignment
- Exact match
- Motif matching
- Dynamic Programming
- String grammar patterns
- Hidden Markov Models
- Phylogeny on sequences

- Graph matching
- Subgraph matching
- Subgraph isomorphism
- Subgraph homorphism
- Approx. graph matching
- Graph grammar patterns
- Graph grammar HMMs ?
- Phylogeny on pathways

Copyright 2002 UC Regents

Protein Interaction Network Databases

- BIND = Biomolecular Interaction DB
- DIP = DB of Interacting Proteins

- These are the two most important, there are others.

BIND=Biomolecular Interaction DB

- **Protein interaction network DB**
- **Pairwise protein interactions = undirected graph**
- **Also other types of data (reactions, …)**
- **http://www.bind.ca**
- **PI: Chris Hogue, Univ. of Toronto**
- **6K interactions, 850 complexes**
- **Includes lots of yeast PI network data**
- **From Yeast 2 Hybrid, and mass spectroscopy expts**
- **ASN.1 import/export**

DIP=Database of Interacting Proteins

- **Protein interactions from experiments**
- **18K protein-protein interactions from 21K experiments**
- **PIs: Ioannis Xenarios and David Eisenberg (UCLA)**
- **Curated**
- **http://dip.doe-mbi.ucla.edu**
- **Available as XML file**
- **Records expt technique, xref to Swiss-Prot, Genbank, PIR**
- **Records binary protein-protein interactions**
- **Graph visualization tool**

Copyright 2002 UC Regents

Survey Paper on Protein Interaction Databases

- Xenarios, I., and D. Eisenberg, "Protein Interaction Databases", *Current Opinions in Biotechnology,* vol. 12, pp. 334-339

Copyright 2002 UC Regents

Biopathways Databases

- Biocyc: Ecocyc, …
- KEGG
- EMP, WIT, …
- Klotho
- aMAZE
- BGDM - Biopathways Graph Data Manager

Copyright 2002 UC Regents

Biocyc

- Ecocyc, Metacyc, ….
- Developed by Peter Karp (SRI), et al.
- Frame representation, Lisp implementation
- Backend = Oracle, frames=blobs
- A dozen organism groups now use
- Data entry, DB, query, graph drawing
- Primarily metabolic pathways, some signaling
- Separate DB for each organism

Copyright 2002 UC Regents

Biocyc (cont.)

- Used for E. coli (Ecocyc)
- Extended to other microbes (Metacyc)
- Query by pathway, EC number (of reaction), reactants, citations, …
- Browse ontologies for reactants, enzymes, reactions, ..
- Applied to several other organisms
- http://www.biocyc.org

Biocyc - complex queries

- Biocyc approach to complex queries
  - Read DB into main memory (in Lisp)
  - Write Lisp program for query
  - Run Lisp program on main memory DB

KEGG

- Kyoto Encyclopedia of Genes and Genomes
- PI: Minoru Kanehisa
- Single composite DB for many organisms
- DB, query facilities, pathways drawings
- URL: http://www.genome.ad.jp/kegg
- Select pathway by:
  - EC number, compound number, gene names

Copyright 2002 UC Regents

EMP, WIT, WIT2

- Eugeni Selkov, Eugeni Selkov, Jr., et al.
- Developed orginally in Russia
- Now at Argonne and Integrated Genomics
- Query pathway by substrate, enzyme, end product, ...
- Servers at IG, Argonne, …
- http://wit.mcs.anl.gov/WIT2

Copyright 2002 UC Regents

Klotho: Biochemical Compunds Declarative Database

- Developed by Toni Kazic (U. Missouri at Columbia)
- Chemical structure graphs of reactants
- Detailed modeling of chemical reaction mechanisms (cf. stoichiometry only in other DB)
- Written in Prolog
- Public server, DB, open source
- http://www.biocheminfo.org/klotho

Copyright 2002 UC Regents

aMAZE

- Metabolic + regulatory pathways database
- Developed by EBI in England
- Shoshana Wodak is PI
- Object Oriented DB
- Entity-association model
- Entities: metabolites, proteins, genes, …
- Associations: reactions, catalysis, transport
- 3 tier: presentation, application, storage

Copyright 2002 UC Regents

aMaze Queries

- Paths: find all paths from A to B
- Pattern search:
  - branch points, feedback loops,
  - pathways affected by a transcription factor
- Pattern discovery (?)
- High level abstraction (?)

Copyright 2002 UC Regents

Biopathways Graph Data Manager

- New project at LBNL
- Graph-based data manager
- Graph data model
- Graph queries
- Funded by DOE GTL and DARPA Biospice
- URL:
  - **http://www.lbl.gov/~olken/graphdm/graphdm.htm**

Copyright 2002 UC Regents

Biopathways Graph Data Model

- Nodes:
  - Chemical entities: metabolites, enzymes, …
  - Bioprocesses: reactions, gene expression, …
- Edges (directed)
  - Indicate relationships
    - input, output, catalyze, inhibit, promote
    - is-a, part-of, element-of, ...

Copyright 2002 UC Regents

Taxonomies

- Of enzymes, reactants, organisms, …
- Simplest are hierarchies (trees)
  - each node has exactly one parent (except root)
  - like library classification systems
- Realistic taxonomies are often DAGs
  - directed acyclic graphs (no cycles)
  - nodes may have multiple parents
- Taxonomies specify partial orders

Copyright 2002 UC Regents

Role of Taxonomies

- Organism, enzyme, metabolite, reaction taxonomies
- Graphs are DAGs (directed acyclic graphs)
- Directed edges = is-a, instance-of
- Node labels are terms
- Terms are used to label nodes in query subgraphs
- Generic terms (upper levels of taxonomy) in query subgraphs must be expanded before performing subgraph matching.
- Example: find reaction containing a kinase enzyme

Why are taxonomies DAGs?

- Simple taxonomies are trees …
- However, some enzymes catalyze more than one kind of chemical reactions
- Hence, some enzymes have more than one parent ==> DAG not tree
- Cycles are forbidden in taxonomies

Graph Queries

- Paths
- k Shortest Paths
- Graph intersection, union, composition
- Graph Matching
  - subgraph isomorphism
  - subgraph homomorphism
  - subgraph homeomorphism
  - approximate graph matching

Copyright 2002 UC Regents

More Graph Queries

- Transitive Closure
- Least Common Ancestor
- Largest Common Subgraph

Copyright 2002 UC Regents

Paths

- Path = connected linear graph
- Example: A--->B--->C--->D
- Length(path) =
  - sum of "lengths of edges" along path
  - typically length of edge = 1

Copyright 2002 UC Regents

Path Queries

- Fixed length patterns
  - commonplace in Object oriented DB, XML DB
- Regular expressions on paths
  - matching labels on nodes (and edges)
  - recursive, arbitrary length paths
  - see work of Mendelzon, etc.
- Shortest path queries
  - k-shortest paths used a surrogate for most important paths in pathways DB
  - well known algorithms

Copyright 2002 UC Regents

Neighborhood Queries

- Neighborhood of radius "r" of subgraph SG of graph G
- Subgraph H of G such that every node in H is within distance "r" of subgraph SG
- distance "r" = length of shortest path
  - edge lengths = 1
- Effect is to include portion of G which is near subgraph SG

Copyright 2002 UC Regents

(Sub)graph Matching

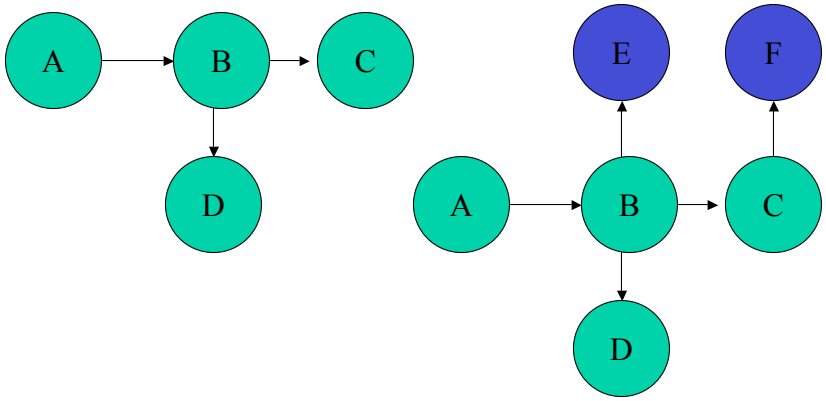
- SG isomorphism
  - exact matching of structure and isomorphism on labels (e.g., match labels also if present)
- SG homorphism
  - exact match of structure
  - labels of query graph nodes are generic terms - require expansion via taxonomy graph
- SG homeomorphism
  - SG homomorphism + ellision of some edges

Copyright 2002 UC Regents

Subgraph isomorphism queries

- Exact match of subgraph structure and labels (sometimes done w/o labels)
- Labels make it easier
- Very common in chemical info retrieval

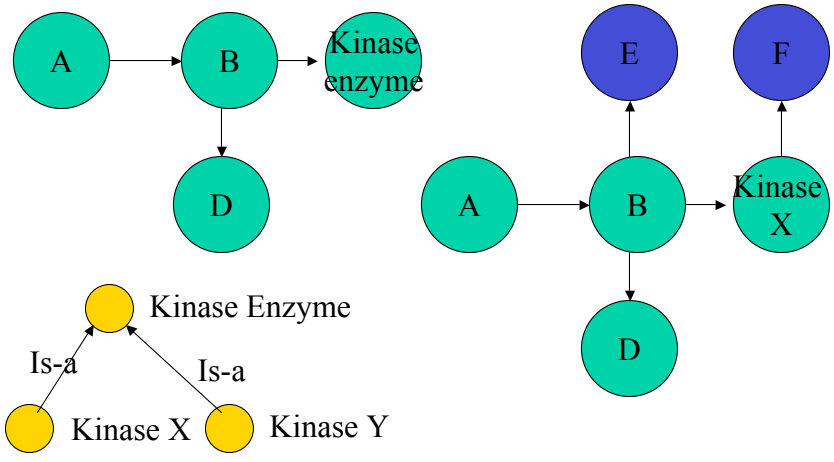
Example subgraph isomorphism



Subgraph homomorphism queries

- Exact match structure (edges)
- Labels on query subgraph are generic terms, e.g., kinase enzyme
- Matching labels are more specific terms subsumed by query node labels, e.g., particular kinase enzymes

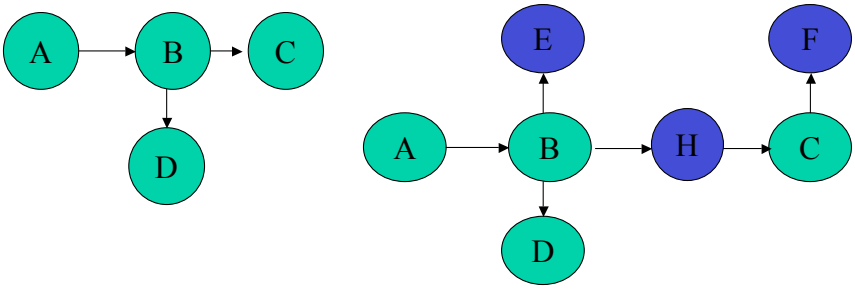
Example subgraph homomorphism



Subgraph homeomorphism

- Match a subgraph G in a graph H, by performing a subgraph isomorphism test against a "contraction of H"
- Contraction of H = contraction of some edge disjoint paths to single edges

Example subgraph homeomorphism



Approximate Graph Matching

- See work by Dennis Shasha, et al.
- Akin to approximate string matching
- Allow: insertion, deletion, substitution of
  - nodes, edges, subgraphs
  - cost for each change
- Dynamic Programming used to find min. cost transformation from graph A to graph B

Node matching

- Needed for (sub)graph matching
- Node match via:
  - exact match of name
  - graph isomorphism of chemical structure graph associated with node
  - approx. string match of sequence (protein, DNA, …)
  - precomputed bipartite matching graph among nodes
  - various algorithmic definitions
  - match node labels + match surrounding context

Boolean Graph Queries

- Graph intersection, union, difference
- Take intersection, union, difference, … of node sets, edge sets
- Note: graph intersection and union can be used to construct majority voting over 3 graphs
- Application: find the difference in metabolisms between two microbes

Majority Graph Query

- Majority (A, B, C) =
- (A&B) U (A&C) U (B&C)
- where
  - & = graph intersection
  - U = graph union
- Usage: to combine multiple (unreliable) protein interaction graphs
- Can be extended to other voting queries

Graph Composition

- Compose two graphs A and B
- Connect outputs of graph A to inputs of graph B
- Used to construct pathways from individual reactions
- Also used to connect pathways, metabolism of co-existing organisms, ...

Shortest Path Queries

- Identify "important" paths from
  - nutrients, or intermediate products
  - to chemical outputs
- Shortest paths queries are attempt to generate most important pathways

Transitive Closure Queries

- Find all products ultimately derived from a particular reaction
- These are potentially affected by knockout (or defect) of root gene
- However, if other paths affect these reactions, then knockout may not inhibit reaction

Least Common Ancestor Queries

- Find the closest ancestor common to sereval nodes in a directed graph
- Observe multiple products are co-regulated
- Identify putative master control reaction
- Classically defined on trees (or DAGs)

Architecture of BDGM

- 3 processes
  - applications, graph data manager, DBMS
- Persistent store = relational DBMS
- Graph query processing in main memory
- Applications programs invoke BGDM via
  - SOAP, XML data exchange
- Applications: pathway viz, editor, analysis
  - (not included)

BGDM cont.

- Just starting up now
- Data sources:
  - **Arkin Lab, VIMSS, Synechococcus, et al.**
  - **Various public biopathways databases**
- BGDM will be open source software
- URL:
  - **http://www.lbl.gov/~olken/graphdm/graphdm.htm**

Network Characterization

- Simple Random Graphs
  - **add edges at random**
  - **degree distribution of nodes is Poisson**
  - **Diameter of graph ~ log(# nodes)**
- Scale-free Graphs
  - **degree distribution satisfies power law**
  - **N(k) ~ k**(-2.5), k=node degree, N(k) = count**
  - **graph diameter = log(log(N)), N=# nodes**

Biopathways Graphs

- **Scale free = power law for node degrees**
- **Self-similar - fractal behavior**
  - **remove some nodes, still have power law distribution of node degrees [Gomez, Lo, Rzhetsky 2001]**
- **Hence, E[diameter] = log(log(N))**
- **Diameter(graph) = longest(shortest path)**
- **Gamma = 2.3 = exponent in power law**

Scale Free Networks

- **Observed for:**
  - **Biopathways**
  - **Web connectivity graphs**
  - **Social networks**
- **Significance**
  - **Much smaller graph diameter**
  - **Graph diameter =**
    - **O(log(log(n)) vs. O(log(n)) for random graphs**
  - **Constrains models of graph evolution**

Conclusions

- **Several biopathways and protein interaction networks DB exist**
- **Very useful for biological research**
- **Typically based on graph data model**
- **These are scale free, self-similar graphs**
- **Presently, limited query facilities**
- **Better graph query capabilities coming**

Contact Information

- Frank Olken
  - **Lawrence Berkeley National Laboratory**
  - **Computational Sciences Research Div.**
  - **1 Cyclotron Road, MS 50B3238**
  - **Berkeley, CA 94720-8147**
  - **http://www.lbl.gov/~olken**
  - **Email: olken@lbl.gov**
  - **Tel: 510-486-5891**

Funding Acknowledgements

- **VIMSS (LBNL Genome to Life Project)**
  - **funded by Dept. of Energy, OBER**
  - **A. Arkin is PI**
- **Synechococcus (Sandia GTL Project)**
  - **funded by DOE OBER, OASCR**
  - **G. Heffelfinger PI (A. Shoshani LBNL PI)**
- **Berkeley BIOSPICE**
  - **funded by DARPA**
  - **A. Arkin is PI**

Personal Acknowledgements

- **Manfred Zorn (LBNL) - assist with figures**
- **Jean Faulon (Sandia) - Network characterization**
- **Sylvia Spengler (NSF) - encouraged work on graph databases**
- **Adam Arkin (LBNL) - supports graph DB work**
- **M. Fernandez, J. Simeone, P. Wadler - Xquery, functional programming approaches to query languages**

Copyright 2002 UC Regents