An Introduction to Pathway Bioinformatics

Yuanhua Tom Tang, Ph.D.
Bioinformatics R & D
Hyseq Pharmaceuticals, Inc.
Sunnyvale, CA, USA

Singapore National University
January, 2002
Definition of Bioinformatics

• Theoretical

The essence of life is information.

Bioinformatics is the study of the information content of life.

• Practical

The essential tool is computer.

Bioinformatics is computer-based information abstraction and processing of biological knowledge.
Pathways

- A schematic diagram of a protein-protein or protein-molecule interaction pathway

A circle indicates a protein or a non-protein biomolecule. An arrow indicates the direction of protein-protein interaction or protein-molecule interaction.
Pathway Database

--- Increasing Level of Complexity

- **The genome**
  - 4 bases
  - 3 billion bp total
  - 3 billion bp/cell, identical

- **The proteome**
  - 20 amino acids
  - ~60K genes, ~200K proteins
  - ~10K proteins/cell; different cells/conditions, different expressions

- **The pathome**
  - ~200K reactions
  - ~20K pathways
  - ~1K pathways/cell; different cells/conditions, different expressions
The Need for Pathway Informatics

• Good angle for data integration and representation.
• Research tool for scientists. Learning tool for students.
• Pharmaceutical drug discovery efforts would benefit from comprehensive pathway databases and tools.
• A challenge for post-genomic era
## List of Pathway Databases/Tools

<table>
<thead>
<tr>
<th>Name</th>
<th>Description</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>KEGG</strong></td>
<td>KEGG is an effort to computerize current knowledge of molecular and cellular biology in terms of the information pathways that consist of interacting molecules or genes and to provide links from the gene catalogs produced by genome sequencing projects. The KEGG project is undertaken in the Bioinformatics Center, Institute for Chemical Research, Kyoto Univ.</td>
</tr>
<tr>
<td><strong>PathDB</strong></td>
<td>PathDB™ is a functional prototype research tool for biochemistry and functional genomics. One of the key underlying philosophies of their project is to capture discrete metabolic steps. This allows them to build tools to construct metabolic networks <em>de novo</em> from a set of defined steps. PathDB is not simply a data repository but a system around which tools can be created for building, visualizing, and comparing metabolic networks.</td>
</tr>
<tr>
<td><strong>Web</strong></td>
<td><a href="http://www.genome.ad.jp/kegg/">http://www.genome.ad.jp/kegg/</a></td>
</tr>
<tr>
<td><strong>Owner</strong></td>
<td>Institute for Chemical Research, Kyoto University</td>
</tr>
<tr>
<td><strong>Web</strong></td>
<td><a href="http://www.ncgr.org/pathdb/index.html">http://www.ncgr.org/pathdb/index.html</a></td>
</tr>
<tr>
<td><strong>Owner</strong></td>
<td>National Center for Genomic Resources</td>
</tr>
</tbody>
</table>
List of Pathway Database/Tools (cont.)

Name:  GenMapp(Gene MicroArray Pathway Profiler)
Gladstone Institute, UCSF.
GenMAPP is a computer application designed to visualize gene expression data on maps representing biological pathways and groupings of genes. The first release of GenMAPP 1.0 beta is available with over 50 mouse and human pathways. They also provide hundreds of functional groupings of genes derived from the Gene Ontology Project for the human, mouse, Drosophila, C. elegans, and yeast genomes. GenMAPP seeks collaborators in the biological community to assist in the development of a library of pathways that will encompass all known genes in the major model organisms.

Name:  SPAD: Signaling Pathway Database
Graduate School of Genetic Resources Technology. Kyushu University.
There are multiple signal transduction pathways: cascade of information from plasma membrane to nucleus in response to an extracellular stimulus in living organisms. Extracellular signal molecule binds specific intracellular receptor, and initiates the signaling pathway. Now, there is a large amount of information about the signaling pathways which control the gene expression and cellular proliferation. They have developed an integrated database SPAD to understand the overview of signaling transduction. SPAD is divided to four categories based on extracellular signal molecules (Growth factor, Cytokine, and Hormone) that initiate the intracellular signaling pathway. SPAD is compiled in order to describe information on interaction between protein and protein, protein and DNA as well as information on sequences of DNA and proteins.
Specific Pathway Databases

- **Cytokine Signaling Pathway DB.** Dept. of Biochemistry. Kumamoto Univ.
  - The Database contains information on signaling pathways of cytokines. It is designed for researchers who work with cytokines and their receptors, and provides biochemical data and references about signaling molecules as well as ligand-receptor relationships.

- **EcoCyc and MetaCyc** Stanford Research Institute
  - EcoCyc database describes the genome and the biochemical machinery of *E. coli*. The database contains up-to-date annotations of all *E. coli* genes. EcoCyc describes all known pathways of *E. coli* small-molecule metabolism. Each pathway and its component reactions and enzymes are annotated in rich detail, with extensive references to the biomedical literature. The Pathway Tools software provides query and visualization services.

**BIND (Biomolecular Interaction Network Database)** UBC, Univ. of Toronto
- BIND is a database designed to store full descriptions of interactions, molecular complexes and pathways, including interactions between any two molecules composed of proteins, nucleic acids and small molecules. Chemical reactions, photochemical activation and conformational changes can also be described. Abstraction is made in such a way that graph theory methods may be applied for data mining. The database can be used to study networks of interactions, to map pathways across taxonomic branches and to generate information for kinetic simulations.
Industrial Companies in Path Informatics

- Protein Pathways, Los Angeles, USA
- Genmetrics, Inc., Silicon Valley, USA
- Biobase, Braunschweig, Germany
- InforMax, Bethesda, MD and AxCell Bioscience, Newtown, PA
- Myriad Proteomics, Salt Lake City, Utah
- CuraGen Corporation, New Haven, CT, USA
Objectives of the KEGG Project

• **Pathway Database**: Computerize current knowledge of molecular and cellular biology in terms of the pathway of interacting molecules or genes.

• **Genes Database**: Maintain gene catalogs of all sequenced organisms and link each gene product to a pathway component.

• **Ligand Database**: Organize a database of all chemical compounds in living cells and link each compound to a pathway component.

• **Pathway Tools**: Develop new bioinformatics technologies for functional genomics, such as pathway comparison, pathway reconstruction, and pathway design.

• **Professor M. Kanehisa** is the leading scientist on the project.
Data Representation in KEGG

- **Entity:** a molecule or a gene
- **Binary relation:** a relation between two entities
- **Network:** a graph formed from a set of related entities
- **Pathway:** metabolic pathway or regulatory pathway
Drosophila melanogaster Genes

According to the KEGG metabolic and regulatory pathways

Pathway Search by [ EC | Cpd | Gene | Seq ]
[ 1st Level | 2nd Level | 3rd Level | Text Search ]

1. Carbohydrate Metabolism
2. Energy Metabolism
   2.1 Oxidative phosphorylation [PATH:dme00190]
   2.2 ATP Synthesis [PATH:dme00193]
   2.4 Carbon fixation [PATH:dme00710]
   2.5 Reductive carboxylate cycle (CO2 fixation) [PATH:dme00720]
   2.6 Methane metabolism [PATH:dme00680]
   2.7 Nitrogen metabolism [PATH:dme00910]
   2.8 Sulfur metabolism [PATH:dme00920]
3. Lipid Metabolism
4. Nucleotide Metabolism
5. Amino Acid Metabolism
6. Metabolism of Other Amino Acids
7. Metabolism of Complex Carbohydrates
8. Metabolism of Complex Lipids
9. Metabolism of Cofactors and Vitamins
Introduction to GenMAPP

- *Gene MicroArray Pathway Profiler* by Bruce Conklin at Gladstone Institute, UCSF.

- GenMAPP is a free computer application designed to visualize gene expression data on maps representing biological pathways and groupings of genes.

- The main features underlying GenMAPP version 1.0 are:
  - Draw pathways with easy to use graphics tools
  - Multiple species gene databases
  - Color genes on MAPP files based on user-imported gene expression data
GenMAPP v1.0 Tutorial
Main Menu

>>> Start Here
Introduction
   What is GenMAPP?
   What kind of data do I need?

Lessons
   1 Drawing a Pathway in GenMAPP
   2 Assigning Gene IDs
   3 Importing a Gene Expression Dataset
   4 Creating Color Sets
   5 Mapping data in GenMAPP
   6 Printing and Exporting to the Web
In Lesson 3 you will be introduced to the Expression Dataset Manager, the interface from which users:

1) Import gene expression datasets and
2) Establish criterion for data visualization

For additional information on this section, click here.
Creating Color Sets

Lesson 4

Gene Color Sets

Criteria Builder

Columns

Name
Control Avg
Exp Avg
Fold Change
P-value

Operations

AND
OR

Label in Legend
New
Save
Add

Once a gene expression dataset has been loaded into the Expression Dataset Manager, a user can begin to establish criteria for visualization in GenMAPP.
Mapping Data in GenMAPP

Once an Expression Dataset has been chosen for any given MAPP file, any other MAPP file you open will automatically have that Color Set applied to it.
Part II. Path Metrics

Software Tools for
Developing Pathway Database,
Performing Pathway Comparison, and
Making Pathway Prediction
Topics to Cover

• SLIPPIR standard for pathway database model
• Gene, pathway, and tissue expression tools
• Pathway search engine
• Ortholog pathway prediction
• Pathway prediction user interface
Path Metrics
PATHWAY SEARCH AND PREDICTION ENGINES

Input Pathway: Input a new pathway into pathway database by providing gene IDs and interactions.

Retrieve Pathway: Retrieve a curated pathway from the database by providing pathway ID or component sequence ID.

Orthologous Pathway: For a given pathway, find orthologous pathway for a given species.

Pathway Search: For a given query pathway, find homologous pathways in the database.

Pathway Prediction: For a given query pathway, predict homologous pathways using gene-gene association data.
SLIPPIR standard for pathway curation

SLIPPIR standards for Standard for LIinear Protein-Protein Interaction Representation.

• For linear comparison (homology),
• 2-D diagrams of pathways ➔ 1-D format.

• We call the 2-D diagrams graph pathways, and the corresponding 1-D pathways linear pathways.

• One graph pathway may be transformed into multiple linear pathways. The generation of graph pathways and the corresponding linear pathways from scientific literature is called pathway curation.

• Pathways are curated by trained scientists with expertise on the relevant pathways. In addition to generating the graph pathway and linear pathways, they also have to generate a pathway description file for each pathway they curate (pathway annotation), and a protein file that contains all the proteins in the pathway.
Mode Symbol Specifications

It is usually specified by two non-character ASCII symbols.

•- > Direct interaction with direction. Used when there is known direct interactions between two nodes (reverse orientation: < -).

•- |  Direct inhibition with direction. Used when there is a direct inhibition from one node to the next. | - for reverse orientation.

•-- Association, indirect action. Used when there is uncertain interaction, indirect interaction, or simply co-expression.

•= = Parallel members. The members can all serve the same function. Usually variants of the same gene, or members from the same family.

•<> Clear interaction, but no direction of information flow (notice, no space within, no letters either). This could happen when more than two proteins are involved to form a large complex.
** Bifurcating members (usually appears only in beginning or ending of a pathway, it can occur in the middle of a pathway only when a pathway bifurcates and immediately folds back, e.g. A->B**C**E->F).

- If a pathway starts to bifurcate in the middle or at the end, one can use a **[path_name] to record this event. E.g:
- A->B->(xx)->C->D**[New_path_1]->E**[New_path_2].

( ) Symbol for non-protein nodes. If the small molecule is uncertain, it can be omitted. If the small molecule is known, its name should be inserted in between, e.g. ->(Ca), or (cAMP).

All the small molecules should be included inside a set of parentheses, e.g. A1->(Ca)->A1->(Cytidine_Diphosphate_Choline).

[ ] Symbol for another pathway. The path_id should be within the bracket.

When linked to other pathways, the path_ids should be put inside a bracket, e.g. A1->[Ca_triggered_path1], A1->[Gs_pathway].

- When an ID is given without a () or [], it means it is a protein node
SLIPPIR Format for Pathway Entries

- The format is based on a common sequence representation format, FASTA
- The pathway will be keyed in FASTA-format, with the top-line being the annotation line. E.g.

```
>PW_ID   PW_name PW_annotation Source Curator Date [Species]
Pr1->Pr2--(Ca)--Pr3==Pr4**Pr5**[PATH_XX]
```

- PW_ID: ID for the pathway
- PW_name: A name
- PW_annotation: a brief description about the pathway
- Source: where this pathway is taken from: article, KEGG, GenMAPP, etc.
- Curator: the person who inputs the pathway
- Date: date of curation
Pathway Database Model (cont.)

- **FASTA format protein-node representation**
  
  `>Seq_id   Annotation
  ABCDELMEN`

  *Comparison Matrix:* percent_identity
  percent_positive (PAM/BLOSSUM)

- **FASTA format non-protein node representation**
  
  `>Mol_id   Annotation
  Molecular structure`

  *Comparison Matrix:* identity mapping
  structural similarity, evolutionary relationship

- **SCOM matrix (similarity coefficient of modes)**
  
  A matrix of numbers, positive and negative values.

  *Comparison Matrix:* identity mapping
  matrix of positive/negative numbers
Pathway Database in Simplest Format

• A SLIPPIR format pathway file
• A FASTA format protein sequence file
• A FASTA format non-protein molecule file
• Flat file tools to do basic database manipulations:
  – Index: generate index file
  – Retrieval: logN scale speed of component access
  – Insertion: cat to the end, new index
  – Deletion: delete, and new index
  – Updating: deletion, cat to the end, new index
Relational Database Implementation
--an example with only protein nodes

Gene_Table
- gene_id
- chromosome
- start
- stop

Protein_Table
- seq_id
- cellular location
- seq_txt
- gene_id (fk)

Interaction_Table
- protein_A
- protein_B
- pathway_id (fk)
- literature_id

Pathway_Table
- pathway_id
- pathway_name
- description
- species
- curator
- entry_data

Protein_Motifs
- motif_id
- seq_id (fk)

Motif_Def_Table
- motif_id
- description
- regular expression
- HMM_matrix

Literature_Table
- literature_id
- author
- journal
- pub_date
- PDF_file

association:
Gene_Table -> Protein_Table: gene_id
Protein_Table -> Interaction_Table: protein=seq_id
Protein_Table -> Pathway_Table: pathway_id
Protein_Motifs -> Motif_Def_Table: motif_id
Protein_Motifs -> Protein_Table: seq_id (fk)
Protein_Motifs -> Literature_Table: literature_id
Pathway_Table -> Literature_Table: literature_id

association (bi-directional):
Gene_Table -> Protein_Table: gene_id
Protein_Table -> Protein_Motifs: seq_id (fk)
Protein_Motifs -> Motif_Def_Table: motif_id
Protein_Motifs -> Protein_Table: seq_id (fk)
Protein_Motifs -> Literature_Table: literature_id
Pathway_Table -> Literature_Table: literature_id
Expression and Expression Comparison

- Gene expression
- Gene expression comparison
- Pathway expression
- Pathway expression comparison
- Tissue expression
- Tissue expression comparison
2. PMsearch Documentation

PMsearch is a pathway comparison program. After a user specifies a query pathway, and a search database, PMsearch will compare the query pathway with each entry in the pathway database. The query pathway is specified by two input files: a query.pw pathway file, and a query.aa, the protein file. The query.pw contains the pathway information, in FASTA format, and the query.aa contains the involved proteins, in FASTA format. The pathway database is also composed of two files, a db.pw and a db.aa file, except the database files contain more than one entry. Once a job is submitted, the search engine (pm_search) will perform the job, and report back all the homologous pathways that are above a user-specified threshold. The user can also specify other parameters, which are given in the user manual.
Given a list of letters, UIPQWEFOIUFJLK and PQEFOIABCDFJ, a good alignment might be:

```
UIPQWXEFOI---UFJLK
|| ||||  ||
PQ--EFOIABCDFJQRS
```

Specifics for pathway alignment:

1. Each letter can represent a node, or a mode.
2. Nodes do not have to be identical in order to match; they just have to be homologous.
3. Distance between nodes and modes, and between protein nodes and non-protein nodes are infinite, you cannot align different types of elements.
In the simplest case, consider pathway with only protein nodes. Given an alignment \( z \), the score is given by

\[
S_z(a,b) = \sum_{t=1}^{k} s(a_{i(t)}, b_{j(t)}) - n_{\text{gap}} \Delta - l_{\text{gap}} \delta
\]

where \( s(x,y) \) is the similarity of protein \( x \) and protein \( y \), \( n_{\text{gap}} \) is the number of gaps in \( z \), \( l_{\text{gap}} \) is the total length of the gaps, \( \Delta \) is a parameter called the “gap opening” penalty, and \( \delta \) is a second parameter called the “gap extension” penalty.

There are many possible alignment for two pathways, and different alignments may have different scores.

PMsearch uses a dynamic programming algorithms to find the alignment with the highest score.
How Alignments Are Determined And Scored

For the alignment to get to \((m,n)\), it must go through one of:

- \((m-1, n-1)\) (\(a_m\) and \(b_n\) are a match),
- \((m-1, n)\) (meaning \((m,n)\) is in a gap in sequence 2),
- \((m, n-1)\) (meaning \((m,n)\) is in a gap in sequence 1).

Recursion:
For \(i = 1\) to \(m\)
  For \(j = 1\) to \(n\)
    \(H(i,j) = \max \{ H(i-1,j-1)+s(i,j), H_h(i,j), H_v(i,j) \} \), where
    \(H_h(i,j) = \max \{ H_h(i,j-1)-d, H(i,j-1)-d-? \} \)
    \(H_v(i,j) = \max \{ H_v(i-1,j)-d, H(i-1,j)-d-? \} \)
  End
End
### PMsearch sample output: list of hits

PMsearch 0.1 Path Metrics [20-Sep-2001] [Build linux x-86 30-Jul-1998]


Query= hsa00625  
(5 proteins)  
PW Database= keggall  
4,881 pathways; 71,600 total proteins.

<table>
<thead>
<tr>
<th>Pathway ID</th>
<th>Pathway Name</th>
<th>Score</th>
</tr>
</thead>
<tbody>
<tr>
<td>hsa00625</td>
<td>Tetrachloroethene degradation</td>
<td>100</td>
</tr>
<tr>
<td>hsa00360</td>
<td>Phenylalanine metabolism</td>
<td>59</td>
</tr>
<tr>
<td>hsa00120</td>
<td>Bile acid biosynthesis</td>
<td>58</td>
</tr>
<tr>
<td>hsa00627</td>
<td>1,4-Dichlorobenzene degradation</td>
<td>40</td>
</tr>
<tr>
<td>hsa00100</td>
<td>Sterol biosynthesis</td>
<td>40</td>
</tr>
<tr>
<td>hsa00940</td>
<td>Flavonoids, stilbene and lignin biosynthesis</td>
<td>40</td>
</tr>
<tr>
<td>hsa00680</td>
<td>Methane metabolism</td>
<td>40</td>
</tr>
<tr>
<td>hsa00950</td>
<td>Alkaloid biosynthesis I</td>
<td>40</td>
</tr>
<tr>
<td>hsa00150</td>
<td>Androgen and estrogen metabolism</td>
<td>40</td>
</tr>
<tr>
<td>hsa00643</td>
<td>Styrene degradation</td>
<td>40</td>
</tr>
<tr>
<td>hsa00380</td>
<td>Tryptophan metabolism</td>
<td>40</td>
</tr>
<tr>
<td>hsa00130</td>
<td>Ubiquinone biosynthesis</td>
<td>40</td>
</tr>
<tr>
<td>hsa00350</td>
<td>Tyrosine metabolism</td>
<td>40</td>
</tr>
<tr>
<td>hsa00340</td>
<td>Histidine metabolism</td>
<td>40</td>
</tr>
<tr>
<td>hsa00053</td>
<td>Ascorbate and aldarate metabolism</td>
<td>28</td>
</tr>
</tbody>
</table>
PMsearch sample output: alignment display

>hsa00340 Histidine metabolism

Query: 4 hsa:51004 hsa:9420 5
%_id: |1.00| |1.00|
Sbjct: 1 hsa:51004 hsa:9420 2

>hsa00053 Ascorbate and aldarate metabolism

Query: 5 hsa:9420 5
%_id: |0.45|
Sbjct: 9 hsa:1582 9

>cel00625 Tetrachloroethene degradation

Query: 1 hsa:51144 hsa:2052 hsa:2053 hsa:51004 4
%_id: |0.39| |0.56| |0.44|
Sbjct: 5 cel:F25G6.5 cel:W01A11.1 --- cel:K07B1.2 7
HOMOLOGS, ORTHOLOGS, AND PARALOGS

Homologs: proteins with good alignment and similar function

Orthologs: proteins performing the same function in different species

Paralogs: homologous proteins in the same species

How to tell the unique ortholog

The ortholog should have a much higher similarity to the query protein that any other protein in its species, and usually higher than most of the paralogs.
EXAMPLE: HOMOLOGS TO THRB_HUMAN

We BLASTed THRB_HUMAN against SwissProt39 and selected the top hits from human and mouse (THRb is the prothrombin precursor). **Orthologs in bold.**

<table>
<thead>
<tr>
<th>HUMAN</th>
<th>MOUSE</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>THRB_HUMAN</strong></td>
<td><strong>THRB_MOUSE</strong></td>
</tr>
<tr>
<td>0.0</td>
<td>2.2e-288</td>
</tr>
<tr>
<td>PRTC_HUMAN</td>
<td>PRTC_MOUSE</td>
</tr>
<tr>
<td>1.3e-61</td>
<td>1.3e-59</td>
</tr>
<tr>
<td>FA10_HUMAN</td>
<td>FA7_MOUSE</td>
</tr>
<tr>
<td>1.4e-54</td>
<td>3.7e-53</td>
</tr>
<tr>
<td>APOA_HUMAN</td>
<td>PLMN_MOUSE</td>
</tr>
<tr>
<td>2.6e-54</td>
<td>1.2e-50</td>
</tr>
<tr>
<td>FA7_HUMAN</td>
<td>HGFL_MOUSE</td>
</tr>
<tr>
<td>3.1e-51</td>
<td>1.4e-40</td>
</tr>
</tbody>
</table>

Note how much higher the similarity is for the ortholog (THRB_MOUSE) whereas the others are in the same range as other paralogs.

**ORTHOLOGOUS PROTEINS OCCUR IN ORTHOLOGOUS PATHWAYS!**
PMortholog Documentation

PMortholog is a simple ortholog prediction program for pathways.

Inputs:
(1) a pathway (query.pw and query.aa files)
(2) a protein database, e.g., SwissProt

• Reports all apparent orthologous pathways

• Most accurate for closely related organisms (e.g. human<->mouse)

• False matches can appear when organisms are too distant, or possibly, because of other paralogous pathways in the organism.
PMOrtholog sample output: hits

PM_ORTHOLOG 0.1, Pathmetrics, Inc. [Oct-20-2001] [Build linux-x86]


Query pathway= hsa00625
(5 proteins)

Database: /u1/pub_db/sp_db/allspecies.aa
374855 proteins.
Summary of ortholog pathways:

<table>
<thead>
<tr>
<th>Hit_nu</th>
<th>species</th>
<th>score</th>
</tr>
</thead>
<tbody>
<tr>
<td>1:</td>
<td>Homo sapiens</td>
<td>100.00</td>
</tr>
<tr>
<td>2:</td>
<td>Mus musculus</td>
<td>65.20</td>
</tr>
<tr>
<td>3:</td>
<td>Rattus norvegicus</td>
<td>65.20</td>
</tr>
<tr>
<td>4:</td>
<td>Caenorhabditis elegans</td>
<td>44.20</td>
</tr>
<tr>
<td>5:</td>
<td>Drosophila melanogaster</td>
<td>37.80</td>
</tr>
<tr>
<td>6:</td>
<td>Arabidopsis thaliana</td>
<td>37.00</td>
</tr>
<tr>
<td>7:</td>
<td></td>
<td>31.80</td>
</tr>
<tr>
<td>8:</td>
<td>Saccharomyces cerevisiae</td>
<td>26.60</td>
</tr>
<tr>
<td>9:</td>
<td>Sinorhizobium meliloti</td>
<td>25.80</td>
</tr>
<tr>
<td>10:</td>
<td>Mesorhizobium loti</td>
<td>24.80</td>
</tr>
<tr>
<td>11:</td>
<td>Agrobacterium tumefaciens</td>
<td>24.80</td>
</tr>
<tr>
<td>12:</td>
<td>Escherichia coli</td>
<td>22.60</td>
</tr>
<tr>
<td>13:</td>
<td>Pseudomonas aeruginosa</td>
<td>22.40</td>
</tr>
<tr>
<td>14:</td>
<td>Schizosaccharomyces pombe</td>
<td>18.80</td>
</tr>
<tr>
<td>15:</td>
<td>Bacillus subtilis</td>
<td>15.00</td>
</tr>
<tr>
<td>16:</td>
<td>Oryza sativa</td>
<td>11.0</td>
</tr>
</tbody>
</table>
PMortholog sample output: alignments

>Hit 1: Ortholog pathway for: Homo sapiens. With score: 100.00
Query:  hsa:51144  hsa:2052  hsa:2053  hsa:51004  hsa:9420
%_id:  1.00  1.00  1.00  1.00  1.00
Sbjct:  gi15082281  gi13097729  gi181395  gi4680659  gi13094303

>Hit 2: Ortholog pathway for: Mus musculus. With score: 65.20
Query:  hsa:51144  hsa:2052  hsa:2053  hsa:51004  hsa:9420
%_id:  0.85  0.88  0.81  0  0.72
Sbjct:  gi3142702  gi12857870  gi12832382  ------  gi12850151

>Hit 3: Ortholog pathway for: Rattus norvegicus. With score: 65.20
Query:  hsa:51144  hsa:2052  hsa:2053  hsa:51004  hsa:9420
%_id:  0.81  0.88  0.84  0  0.73
Sbjct:  gi4098957  gi207689  gi55930  ------  gi1226240

>Hit 4: Ortholog pathway for: Caenorhabditis elegans. With score: 44.20
Query:  hsa:51144  hsa:2052  hsa:2053  hsa:51004  hsa:9420
%_id:  0.48  0.56  0.42  0.44  0.31
Sbjct:  gi726418  gi1465805  gi3876864  gi2088820  gi13775482
#!/usr/bin/perl

# program: pm_ortholog
# purpose: finds an orthologous pathway for a query pathway in a given species. Prints the output in alignment format.
#
# author: Grace Yang
# Pathmetrics, Inc.
# 10/14/2001
#
# usage: pm_ortholog <query_pw> <query_aa> <protein_db>
# were query_path.pw contains the pathway information
# query_path.aa contains all the proteins in query

use strict;

# Part 1. Parse input, check files

my ($usage, $q_id, $q_aa, $q_pnu, $q_pw, $aa_db);
my (%gn2spec, %score, %total_score, $file);
my (@q, @arr, %qu2spec, $spec, @time_st);

$usage = "\n$0 <query_pw> <query_aa> <protein_db>\n    query_pw: query pathway file
    query_aa: query aa file
    protein_db: protein db to search\n\n";

if (@ARGV<1) { die "$usage";}

($q_pw, $q_aa, $aa_db)=@ARGV;
for $file ("$q_pw", "$q_aa", $aa_db) { if (!(-e "$file")) { die "Did not find $file file\n"; }}
open (QSEQ, "$q_pw");
while (<QSEQ>) {
    $file= $_; chomp ($file);
    if ($file=~/^>(\S+)\s/){ $q_id=$1; next;}
    push (@q, split (/\s+/, $file)); $q_pnu= @q;
} close (QSEQ);

@time_st = localtime; &print_header;
&big_matrix_sort ($aa_db, $q_aa);

open (AA, "/usr/local/biobin/im_retrieve $aa_db /tmp/$$.matrix.ids |")
while (<AA>) { if ($_=~/^>(\S+)\s+.*\[(\[\w\s\]+)\]/) { $gn2spec{$1} = $2;}}
close (AA);

# get the best hit for each query id and each spec
open (MAT, "/tmp/$$.matrix.s");
while(<MAT>) {
    chomp; @arr = split (/\t/);
    if ($qu2spec{$arr[0]}->{ $gn2spec{$arr[1]} }) { next; }
    $qu2spec{$arr[0]}->{ $gn2spec{$arr[1]} } = $arr[1];
    $score{$arr[0]}->{ $arr[1] } = $arr[2];
    if ($total_score{$gn2spec{$arr[1]}}){
        $total_score{$gn2spec{$arr[1]}} += $arr[2]*20;
    } else{ $total_score{$gn2spec{$arr[1]}} = $arr[2]*20;
    } close(MAT);
my ($qid, $i, $j, $ln); $ii=0;
foreach $spec (sort by_score keys (%total_score)) { $ii++;
    printf "Hit%3d: Ortholog pathway for: %20s. With score: %5.2f\n\n", $ii,$spec,
    $total_score{$spec};
    for ($i=0; $i<(@q/6); $i++) {
        my (@ln1, @ln2, @ln3, $sc, $hid, $k);
        for ($j=0; $j<6; $j++) {
            $k = $i*6+$j;
            if ($k <@q){ $sc = $score{$q[$k]}{$qu2spec{$q[$k]}->{$spec}};
                if ($qu2spec{$q[$k]}->{$spec}) {$hid=$qu2spec{$q[$k]}->{$spec};
                    } else {$hid ="------";}
                if (!defined($sc)) {$sc=0.0;}
                push (@ln1,$q[$k]);push (@ln2, "\|$sc\|"); push (@ln3, $hid);
            }
        }
    format STDOUT=
    Query: @|||||||||| @|||||||||| @|||||||||| @|||||||||| @|||||||||| @||||||||||
    @|||||||||| $ln1[0], $ln1[1], $ln1[2],$ln1[3],$ln1[4],$ln1[5]
    %_id:      @|||     @|||      @|||       @|||     @|||     @|||     @|||
    @|||||
    $ln2[0], $ln2[1], $ln2[2],$ln2[3],$ln2[4],$ln2[5]
    Sbjct: @|||||||||| @|||||||||| @|||||||||| @|||||||||| @|||||||||| @||||||||||
    @|||||||||| $ln3[0], $ln3[1], $ln3[2],$ln3[3],$ln3[4],$ln3[5]
    .
    write STDOUT; }
}

&print_end;
sub by_score { return $total_score{$b}<=$total_score{$a};}

sub big_matrix_sort {
    my (@arr, $q_len, $m_len, $pct_id, $pct_pos, $l, $tp);
    my ($bg, $end, $hsp_len, $pm_score);
    my ($aa_db, $qu_aa)=@_;  
    open (IN, "'/usr/local/biobin/im_cycle blastp $aa_db $q_aa S=100 | '/usr/local/biobin/pm_pblast | ");
    open(HIT, "'/tmp/$$.matrix');
    while(<IN>){
        chomp; @arr = split(/\t/);
        ($q_len, $m_len) = split(/:/,$arr[2]);
        ($pct_id, $pct_pos) = split(/:/, $arr[5]);
        ($l, $tp) = split(/:/, $arr[6]);
        ($bg, $end) = split(/-/, $l);
        $hsp_len = abs($end-$bg)+1;

        $pm_score = get_pm_score($pct_id, $pct_pos, $hsp_len, $q_len, $m_len);
        if($pm_score <= 0) { next; }  
        printf HIT "%s\t%s\t%3.2f\n", $arr[0],$arr[1],$pm_score;
    }
    close(IN);close(HIT);
    system ("sort -k 3rn /tmp/$$.matrix >/tmp/$$.matrix.s");
    system ("cut -f2 /tmp/$$.matrix |sort -u >/tmp/$$.matrix.ids");
}
sub get_pm_score {
  my ($pct_id, $pct_pos, $hsp_len, $q_len, $m_len) = @_; 
  my $len = ($q_len<$m_len) ? $q_len : $m_len;
  if($len <= 0) {
    #print STDERR "warn: length of sequence is calculated to <= 0\n";
    return -1;
  }else{ return 0.005 * ($pct_id + $pct_pos) * $hsp_len / $len;}
}

sub print_header {
  my ($aa_nu);

  print "\n";
  print "PM_ORTHOLOG 0.1, Pathmetrics, [Oct-20-2001] [Build linux-x86]\n\n";
  print "Ref.: US Pat.Pending. "Methods for Establishing Pathway Database\n";
  print "and Perform Pathway Searches". XXX Feb. 20, 2001.\n\n";

  print "Query pathway= $q_id\n";
  print " ($q_pnu proteins)\n\n";
  print "Database: $aa_db\n";
  open (DB, "$aa_db.db");
  while (<DB>) {if ($_=~/Total keys\s(\d+)/) {$aa_nu=$1; last;}}
  close (DB);
  print " $aa_nu proteins.\n";
}
Pathway Prediction Engines

• They are the crown jewels of Pathmetrics software tools
• Can predict many novel interactions
• Use diverse input data, including sequence data, expression data, and known interaction data
• Employ complex numerical algorithms such as dynamical programming and clustering
Pathway Prediction

Please specify your query: (input 1 or 2)

1. Query Path_id
2. Query <.pw> file

Please specify gene-gene association database:

Additional options:

- Homologous pathway prediction
- Reporting p_value
- Suboptimal path to report
- Hits to show in summary line
- Hits to show in alignment

Submit Clear
### Prediction Searching Result

**Query path:**

```plaintext
gi14448283 gi18169051 gi114741745 gi13540998 gi1736683 gi136454
gi181976 gi6164681 gi190918 gi496891 gi1777994 gi101155
gi187717 gi13748677 gi126284 gi182735 gi1386839
```

PM_PREDICT 0.1, Pathmetrics, Inc. [Oct-20-2001] [Build linux-x86]


Application number 60/269,711.

**Query pathway**: EPO

(17 proteins)

**Database**: /u1/pub_db/Tom.asm

15903 proteins.

**Summary of predicted pathways:**

<table>
<thead>
<tr>
<th>Putative pathway ID</th>
<th>status</th>
<th>score</th>
</tr>
</thead>
<tbody>
<tr>
<td>putative_pw_1</td>
<td>optimal</td>
<td>87.35</td>
</tr>
<tr>
<td>putative_pw_2</td>
<td>optimal</td>
<td>81.60</td>
</tr>
<tr>
<td>putative_pw_3</td>
<td>optimal</td>
<td>79.95</td>
</tr>
<tr>
<td>putative_pw_4</td>
<td>optimal</td>
<td>75.95</td>
</tr>
<tr>
<td>putative_pw_5</td>
<td>sub-optimal</td>
<td>81.60</td>
</tr>
<tr>
<td>putative_pw_6</td>
<td>sub-optimal</td>
<td>79.95</td>
</tr>
<tr>
<td>putative_pw_7</td>
<td>sub-optimal</td>
<td>75.95</td>
</tr>
</tbody>
</table>

>putative_pw_1 Optimal predicted pathway with score: 87.35
<table>
<thead>
<tr>
<th>Input Pathway</th>
<th>Retrieve Pathway</th>
<th>Pathway Search</th>
<th>Orthologous Pathway</th>
<th>Pathway Prediction</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>&gt;putative_pw_1</strong></td>
<td>Optimal predicted pathway with score: 87.35</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Query:</td>
<td>gi14741745</td>
<td>1</td>
<td>gi12560928</td>
<td>1</td>
</tr>
<tr>
<td>% id</td>
<td>[10.42]</td>
<td></td>
<td>[10.41]</td>
<td></td>
</tr>
<tr>
<td>Sbjct</td>
<td>Hs.50651</td>
<td>0.49</td>
<td>Hs.75648</td>
<td>0.38</td>
</tr>
<tr>
<td>Query:</td>
<td>gi36454</td>
<td>1</td>
<td>gi181976</td>
<td>1</td>
</tr>
<tr>
<td>% id</td>
<td>[11.00]</td>
<td></td>
<td>[11.00]</td>
<td></td>
</tr>
<tr>
<td>Sbjct</td>
<td>Hs.81972</td>
<td>0.51</td>
<td>Hs.296381</td>
<td>0.38</td>
</tr>
<tr>
<td>Query:</td>
<td>gi190210</td>
<td>1</td>
<td>gi496091</td>
<td>1</td>
</tr>
<tr>
<td>% id</td>
<td></td>
<td>[0.99]</td>
<td></td>
<td>[0.99]</td>
</tr>
<tr>
<td>Sbjct</td>
<td>Hs.184050</td>
<td>0.58</td>
<td>Hs.85181</td>
<td>0.69</td>
</tr>
<tr>
<td>Query:</td>
<td>gi181155</td>
<td>1</td>
<td>gi187517</td>
<td>1</td>
</tr>
<tr>
<td>% id</td>
<td></td>
<td>[10.99]</td>
<td></td>
<td>[10.53]</td>
</tr>
<tr>
<td>Sbjct</td>
<td>Hs.165843</td>
<td>0.33</td>
<td>Hs.9302</td>
<td>0.31</td>
</tr>
<tr>
<td>Query:</td>
<td>gi4126584</td>
<td>1</td>
<td>gi182735</td>
<td>1</td>
</tr>
<tr>
<td>% id</td>
<td></td>
<td>[1.00]</td>
<td></td>
<td>[1.00]</td>
</tr>
<tr>
<td>Sbjct</td>
<td>Hs.25647</td>
<td>0.65</td>
<td>Hs.78465</td>
<td>14</td>
</tr>
<tr>
<td><strong>&gt;putative_pw_2</strong></td>
<td>Optimal predicted pathway with score: 18.45</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Query:</td>
<td>gi187517</td>
<td>1</td>
<td>gi12548677</td>
<td>1</td>
</tr>
<tr>
<td>% id</td>
<td></td>
<td>[10.34]</td>
<td></td>
<td>[10.20]</td>
</tr>
<tr>
<td>Sbjct</td>
<td>Hs.94576</td>
<td>0.31</td>
<td>Hs.180383</td>
<td>0.46</td>
</tr>
<tr>
<td>Query:</td>
<td>gi182735</td>
<td>1</td>
<td>gi386839</td>
<td>17</td>
</tr>
<tr>
<td>% id</td>
<td></td>
<td>[10.30]</td>
<td></td>
<td>[0.55]</td>
</tr>
<tr>
<td>Sbjct</td>
<td>Hs.79678</td>
<td>0.41</td>
<td>Hs.198951</td>
<td>4</td>
</tr>
<tr>
<td><strong>&gt;putative_pw_3</strong></td>
<td>Optimal predicted pathway with score: 5.95</td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>