



The MetaCyc Metabolic Pathway Database

Peter D. Karp, Ph.D.
Bioinformatics Research Group
SRI International
pkarp@ai.sri.com

<http://www.ai.sri.com/pkarp/>
<http://MetaCyc.org/>

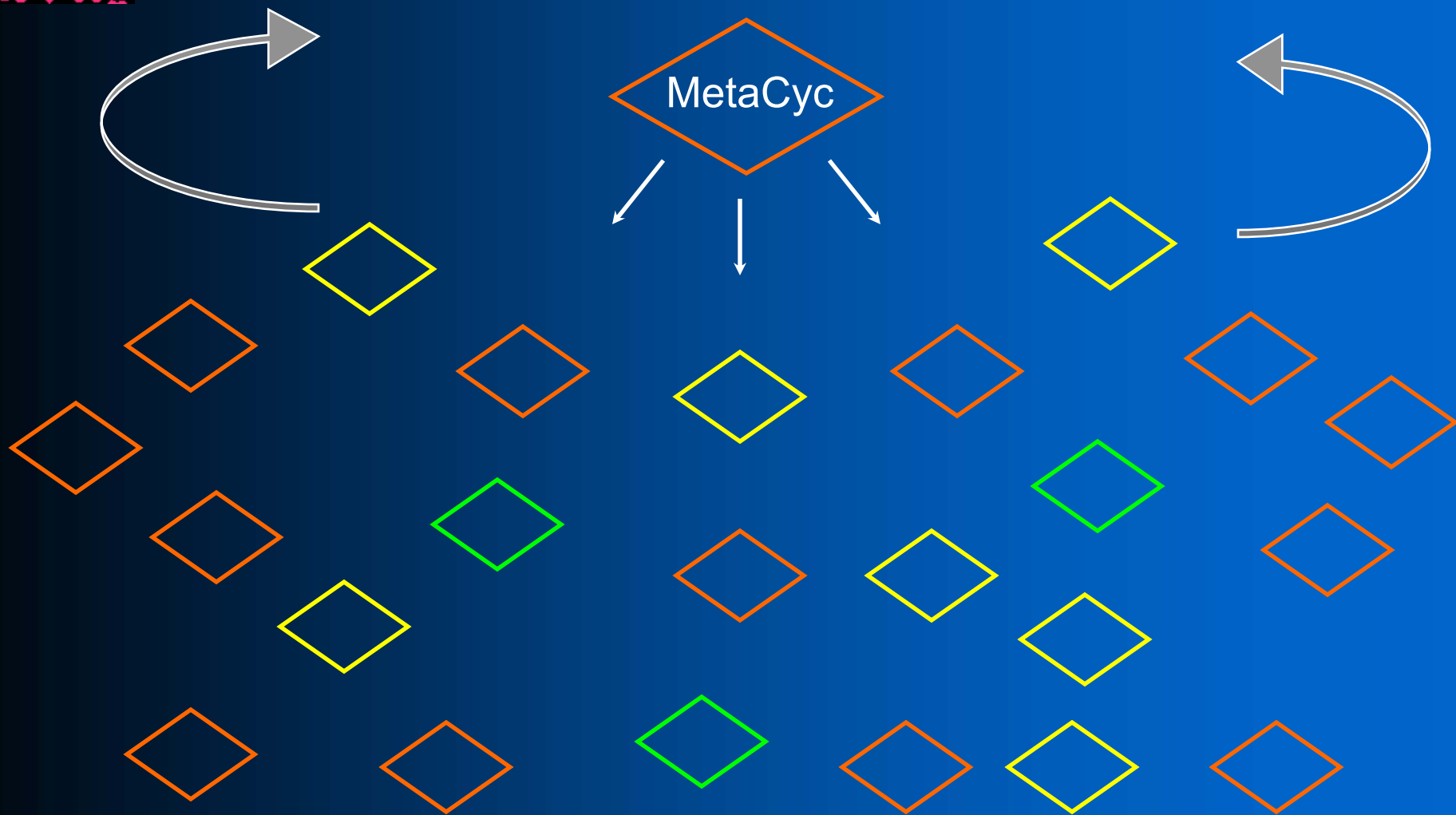
Terminology

- **Model Organism Database (MOD)** – DB describing genome and other information about an organism
- **Pathway/Genome Database (PGDB)** – MOD that combines information about
 - Pathways, reactions, substrates
 - Enzymes, transporters
 - Genes, replicons
 - Transcription factors, promoters, operons, DNA binding sites
- **BioCyc – Collection of PGDBs at BioCyc.org**
 - Organism-specific PGDBs: EcoCyc, PseudoCyc
 - Multi-organism PGDB: MetaCyc

The screenshot shows a Netscape browser window displaying the BioCyc Knowledge Library homepage. The browser's address bar shows the URL <http://biocyc.org/>. The page features a navigation menu on the left with links for Home, Search, News, Services, Information, and Databases. The main content area is titled "BioCyc Home Page" and provides an overview of the BioCyc Knowledge Library as a collection of Pathway/Genome Databases. It lists various databases, including literature-derived ones like EcoCyc and MetaCyc, and computationally-derived ones like AgroCyc, BsubCyc, CtraCyc, CauloCyc, HpvCyc, HlnCyc, MtbRvCyc, MpnCyc, PseudoCyc, YeastCyc, TpalCyc, and VchoCyc. An "Acknowledgments" section at the bottom credits funding agencies like DARPA, the Department of Energy, and the NIH for their support in developing the databases.

Family of Pathway/Genome Databases

SRI International
Bioinformatics



Terminology



- **Pathway Tools Software – Software for generating, curating, querying, displaying PGDBs**
 - PathoLogic – Infer pathways from genome
 - Pathway/Genome Editors – Distributed curation environment
 - Pathway/Genome Navigator – Query, visualization, analysis, Web publishing

Pathway Definition



- A chemical reaction interconverts chemical compounds



- An enzyme is a protein that accelerates chemical reactions

- A pathway is a linked set of reactions



- A conceptual unit of cell's biochemical machine

MetaCyc: Metabolic Encyclopedia

- **Goal: Describe a representative sample of every experimentally determined metabolic pathway**
- **Literature-based DB with extensive references and commentary**
- **Pathways, reactions, enzymes, substrates**
- **460 pathways, 1262 enzymes, 4293 reactions**
 - 172 *E. coli* pathways, 2735 citations
- ***Nucleic Acids Research* 30:59-61 2002.**
- **Jointly developed by SRI and Carnegie Institution**
 - New focus on plant pathways
 - Related PGDB: AraCyc -- see www.arabidopsis.org:1555

MetaCyc Frequent Organisms

<i>E. coli</i>	173
<i>Sm. typhimurium</i>	35
<i>Ho. sapiens</i>	31
<i>Sf. sulfataricus</i>	20
<i>B. subtilis</i>	18
Soybean	18
<i>Pseudomonas</i>	17
<i>Hp. influenzae</i>	15
<i>M. capricolum</i>	12
<i>S. cerevisiae</i>	8
<i>P. putida</i>	7
<i>M. pneumoniae</i>	7

MetaCyc Data



- **MetaCyc contains one DB object for each distinct pathway**
 - Distinct in terms of reaction steps
 - Each pathway labeled with species it occurs in
- **MetaCyc pathways are experimentally determined**
- **4218 reactions in MetaCyc**
 - 401 lack EC numbers

MetaCyc Enzyme Data



- **Reaction(s) catalyzed**
- **Alternative substrates**
- **Cofactors / prosthetic groups**
- **Activators and inhibitors**
- **Subunit structure**
- **Molecular weight, pI**
- **Comment, literature citations**
- **Species**

MetaCyc Pathway Variants



- **Pathways that accomplish similar biochemical functions using different biochemical routes**
 - Alanine biosynthesis I – *E. coli*
 - Alanine biosynthesis II – *H. sapiens*
- **Pathways that accomplish similar biochemical functions using similar sets of reactions**
 - Several variants of TCA Cycle

MetaCyc Super-Pathways



- **Group together sets of pathways that are linked by common substrates**
- **Example: Super-pathway containing**
 - Chorismate biosynthesis
 - Tryptophan biosynthesis
 - Phenylalanine biosynthesis
 - Tyrosine biosynthesis
- **Super-pathways defined by listing their component pathways as components**
- **Multiple levels of super-pathways can be defined**
- **Pathway layout algorithms accommodate super-pathways**

MetaCyc Hierarchical Taxonomies



- **Classification systems for grouping information**
- **Pathways**
- **Compounds**
- **Reactions**

MetaCyc Data Validation

- **Database consistency constraints**
- **Element balancing of reactions**



BioCyc Availability



- **WWW BioCyc**
 - EcoCyc, MetaCyc
 - Pathway/genome DBs for 13 other organisms
- **<http://biocyc.org/>**
- **<http://biocyc.org:1555/server.html>**

- **Downloadable BioCyc**
 - Flatfiles
 - Binary executables: Hardware requirements
 - ◆ Sun UltraSparc-170 w/ 64MB memory
 - ◆ PC, 500MHz CPU, 128MB memory, Windows-98 or newer

Applications of MetaCyc



- **Reference sources on metabolism**
- **Sequence/pathway analysis of microbial genomes**
- **Analysis of gene-expression data**
- **Computer-aided education**
- **Anti-microbial drug discovery**
- **Pathway engineering**
- **Investigations of**
 - Comparative metabolism
 - Global properties of *E. coli* metabolic network
 - Sequence / structure / pathway relationships

Representation of Metabolic Pathways



- **The pathway graph representation:**
 - Graph nodes for each enzyme and substrate
 - Graph edges connect substrates
- **This representation is complete, but it is redundant with other parts of MetaCyc**
- **Redundancy slows knowledge acquisition, can lead to inconsistencies due to partial updates**
- **Preferred approach: Layer pathways on top of reaction and compound data already in MetaCyc**

Predecessor-List Representation



- Defines pathways by reference to existing reactions
- Use inference to determine reaction directions
- Use automatic layout to derive node positions



Predecessor-List Representation

- A pathway is defined as a list of reaction pairs
- The predecessor-list representation defines a pathway as a set of predecessor/successor relationships among reactions
- **Predecessor list:**
 - 1 precedes 2 ; 2 precedes 3

1. $A \longrightarrow B$

2. $C \longrightarrow B$

3. $C \longrightarrow D$

$A \longrightarrow B \longrightarrow C \longrightarrow D$

Pathway Display



- **Navigator generates textbook-style layouts of biochemical pathways**
- **Display algorithm:**
 - Determine direction for each reaction
 - Determine main and side compounds
 - Construct pathway graph
 - Determine topology of pathway graph
 - Apply layout algorithm to pathway graph
 - ◆ Linear
 - ◆ Tidy tree
 - ◆ Circular
 - ◆ Combination

Pathway Tools Software



- **PathoLogic**
 - Prediction of metabolic network from genome
 - Computational creation of new Pathway/Genome Databases
- **Pathway/Genome Editors**
 - Distributed curation of genome annotations
 - Distributed object database system
 - Interactive editing tools
- **Pathway/Genome Navigator**
 - WWW publishing of PGDBs
 - Graphic depictions of pathways, chromosomes, operons
 - Analysis operations
 - ◆ Pathway visualization of gene-expression data
 - ◆ Global comparisons of metabolic networks

Pathway/Genome Navigator



- **Algorithmic visualization of pathway and genome data**
- **Predefined queries for each object type**
 - Get pathway by name, substring, class, species
- **X-windows and WWW**
- **PathoLogic and Editors run through X-windows only**

Visualization and Editing Tools



- **Full Metabolic Map**
- **Pathways**
- **Reactions**
- **Compounds**
- **Enzymes, Transporters, Transcription Factors**

- **Genes**
- **Chromosomes**
- **Operons**

Inference of Metabolic Pathways

SRI International
Bioinformatics



ANNOTATED GENOME
Structured ASCII Text File

MetaCyc

PathoLogic

Pathway/Genome Database

List of Gene Products

List of Genes/ORFs

DNA Sequence

Pathways

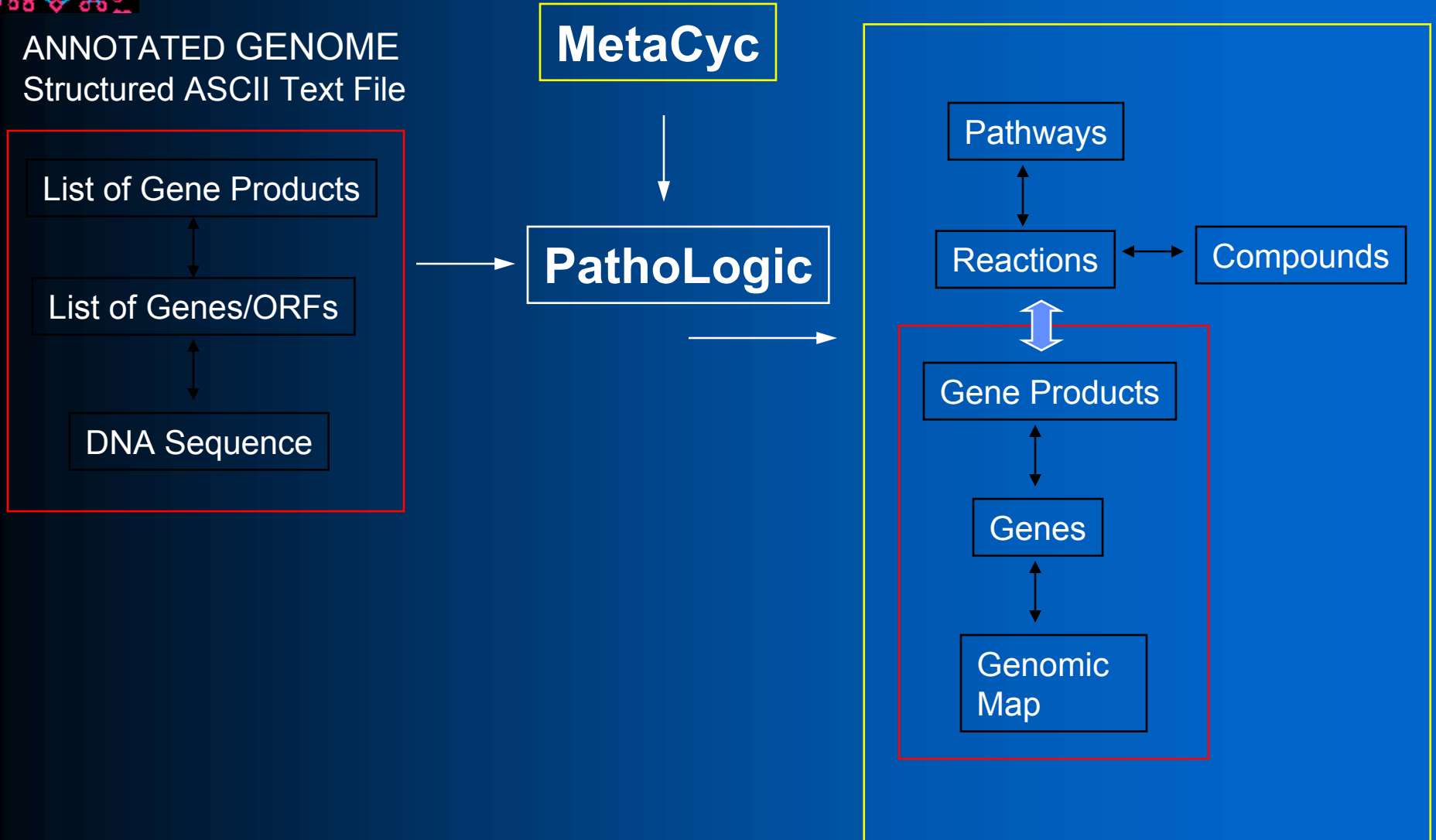
Reactions

Compounds

Gene Products

Genes

Genomic
Map



Genbank Format :

```
gene          422054..423490
              /gene="aroE"
CDS           422054..423490
              /gene="aroE"
              /label="CT370"
              /product="Shikimate 5-Dehydrogenase"
              /db_xref="PID:g3328794"
```

PathoLogic Format :

```
ID           CT370
NAME         aroE
STARTBASE   422054
ENDBASE     423490
PRODUCT     Shikimate 5-Dehydrogenase
DBLINK      PID:g3328794
PRODUCT-TYPE P
EC          1.1.1.25
```

PathoLogic Analysis Phases

- **Trial parsing of input data files**
- **Automated build of initial PGDB**
 - Initialize schema of new PGDB
 - Create DB objects for chromosomes, genes, proteins
 - Match proteins to reactions via EC number and enzyme name
 - Import corresponding pathways from MetaCyc
- **Refine protein/reaction mapping**
- **Define protein complexes**
- **Define metabolic overview diagram**

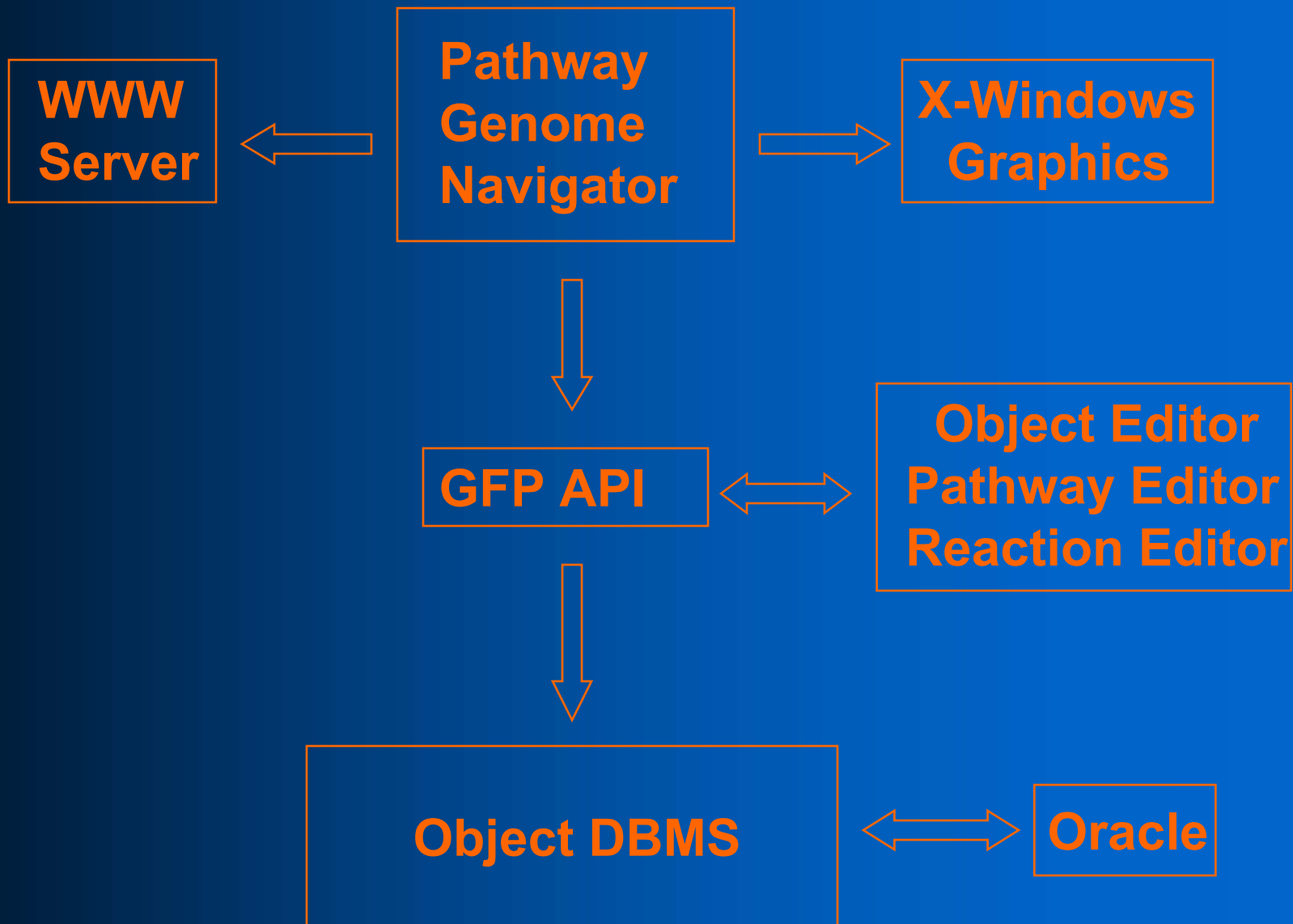




PathoLogic: Inference of Pathway Complement

- **Extends the paradigm of genome analysis**
- **Predicted genes placed in their biochemical context**
 - Information reduction device
 - Assess coherence of the set of genes in a genome
 - Identifies pathway holes and singleton enzymes
 - Provides a framework for analysis of functional-genomics data

Pathway Tools Architecture



Computing with MetaCyc: APIs



- **PerlCyc interface**

- Library of Perl functions for querying MetaCyc and other PGDBs via socket connection
- Database access functions
 - ◆ Select_Organism, All_Pathways
- Functions for performing inference / hardwired queries
 - ◆ Genes_Of_Reaction, Genes_Of_Pathway
 - ◆ Transcription_Unit_Transcription_Factors
 - ◆ Enzyme_P

- **JavaCyc interface also in progress**

- <http://aracyc.stanford.edu/~mueller/perlcyc/>

- **Lisp API**

- <http://bioinformatics.ai.sri.com/ptools/ptools-resources.html>



Computing with MetaCyc: Flat Files

- **Two file formats: tab-delimited, attribute-value**
- **One file for each format, each datatype**
- **Specification:**
 - <http://bioinformatics.ai.sri.com/ptools/flatfile-format.html>
- **Examples:**
 - Pathways.col – Pathways and genes encoding enzymes
 - Enzymes.col – Enzymes and reactions they catalyze
 - Pathways.dat – Full data on each pathway
 - Reactions.dat – Full data on each reaction

Example Flatfile – Pathways.dat




```
UNIQUE-ID - P107-PWY
TYPES - Energy-Metabolism
COMMON-NAME - RuMP cycle and formaldehyde assimilation
REACTION-LIST - FORMATEDEHYDROG-RXN
REACTION-LIST - FORMALDEHYDE-DEHYDROGENASE-RXN
REACTION-LIST - 6PGLUCONDEHYDROG-RXN
REACTION-LIST - R84-RXN
REACTION-LIST - PGLUCISOM-RXN
REACTION-LIST - R12-RXN
REACTION-LIST - R10-RXN
SYNONYMS - ribulose-monophosphate cycle
SYNONYMS - formaldehyde oxidation
//
```

Example Flatfile – Reactions.dat

```
UNIQUE-ID - R84-RXN
TYPES - EC-1.1.1
EC-NUMBER - 1.1.1.-
IN-PATHWAY - P122-PWY
IN-PATHWAY - P107-PWY
LEFT - GLC-6-P
LEFT - NAD
OFFICIAL-EC? - NO
RIGHT - 6-P-GLUCONATE
RIGHT - NADH
RIGHT - PROTON
//
```

Example Flatfile – Compounds.dat



```
UNIQUE-ID - GLC-6-P
TYPES - Carbohydrate-Derivatives
COMMON-NAME - glucose-6-phosphate
CAS-REGISTRY-NUMBERS - 56-73-5
CHEMICAL-FORMULA - (C 6)
CHEMICAL-FORMULA - (H 13)
CHEMICAL-FORMULA - (O 9)
CHEMICAL-FORMULA - (P 1)
MOLECULAR-WEIGHT - 260.137
SYNONYMS - D-glucose-6-P
SYNONYMS - glucose-6-P
SYNONYMS - &alpha;-D-glucose-6-phosphate
SYNONYMS - &alpha;-D-glucose-6-P
SYNONYMS - D-glucose-6-phosphate
//
```

Comparison of MetaCyc to KEGG



• Data

- MetaCyc contains more pathways, more detail
- KEGG has no literature citations, no comments, no detailed information about enzymes (inhibitors, subunits)
- KEGG does not model pathways of any specific organism
- KEGG cannot indicate absence of a pathway, or specify what part of a large map is present in a given organism
- KEGG difficult to compute with

• Software tools

- KEGG has no algorithmic visualization tools
- KEGG has no queryable metabolic-map overview diagram
- KEGG has no interactive editing tools

Summary



- **MetaCyc is a comprehensive metabolic pathway DB**
- **Literature based**
- **Detailed information on each pathway**
- **Goal is to contain an example of every different metabolic pathway**
- **Freely available**

Acknowledgements

● SRI

- Cindy Krieger, Suzanne Paley, John Pick

● Carnegie Institution

- Chris Somerville, Sue Rhee, Lukas Mueller, Peifen Zhang

● John Ingraham

● Funding sources:

- NIH National Institute of General Medical Sciences

pkarp@ai.sri.com

MetaCyc.org