Development of Large Scientific Knowledge Bases

Peter D. Karp, Ph.D.

Bioinformatics Research Group

SRI International

pkarp@ai.sri.com

BioCyc.org

EcoCyc.org, MetaCyc.org, HumanCyc.org

The Problems

- Create browsable online encyclopedias for genomes and cellular networks
- Computational symbolic theories needed as theory complexity increases
- Quickly assemble an organism-specific database
- Refine that database by incorporating information from the biomedical literature
- Enable biologists to issue precise queries to complex databases
- Applications that solve problems in computational biology drawing from these databases

The Solutions

- **The content development approach**
  - Computational inference tools
  - Knowledge entry by PhD-level curators
- **Web query and visualization tools provide scientists with access**
- **An ontology enables high-fidelity knowledge representation**
- **Ocelot KB management environment supports distributed development of large KBs by multiple users**
- **Applications**
- **Common Lisp**

What to do When Theories Become Larger than Minds can Grasp?

- **Example: *E. coli* metabolic network**
  - 180 pathways involving 744 reactions and 791 substrates
- **Example: *E. coli* genetic network**
  - Control by 97 transcription factors of 1174 genes in 630 transcription units

- **Past solutions:**
  - Experts specialize
  - Publish theories in textual form

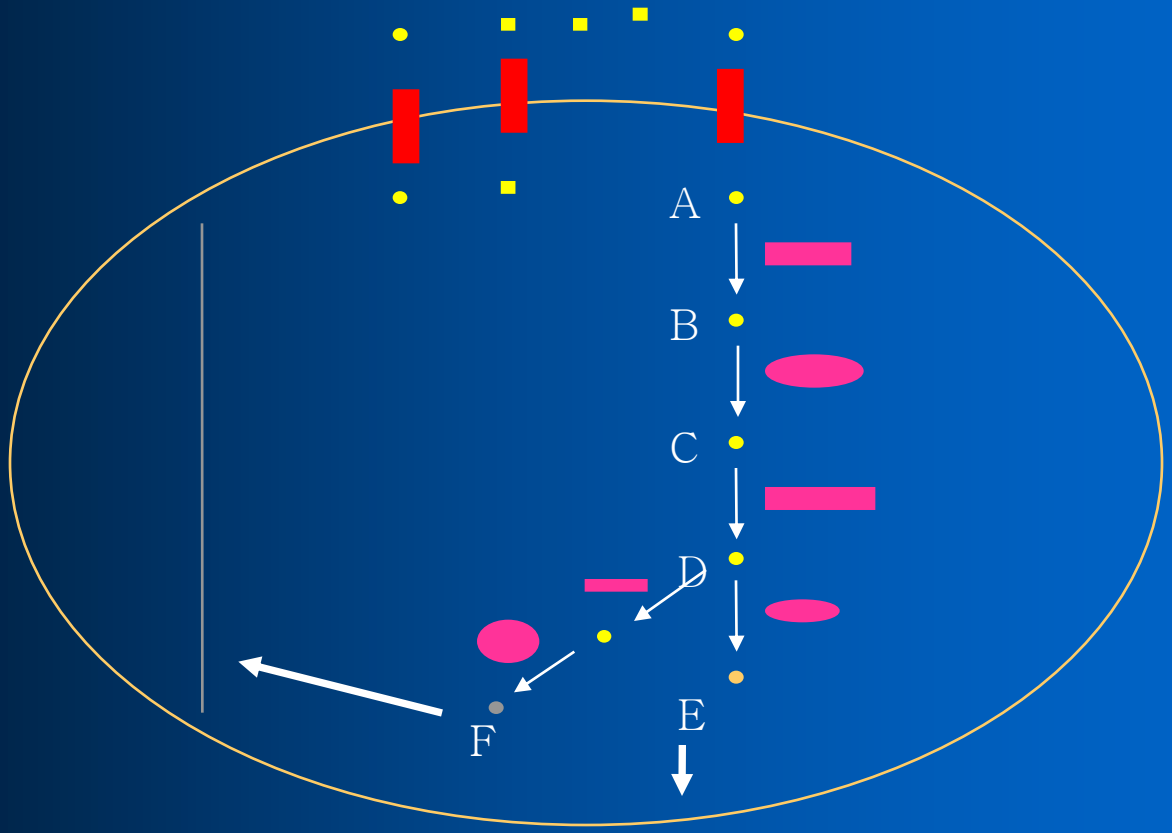
- **We cannot compute with theories in those forms**
  - Evaluate theories for consistency with new data: microarrays
  - Refine theories with respect to new data
  - Compare theories describing different organisms

Biological Knowledge Bases

- **Store biological knowledge and theories in computers *in a declarative form***
  - Accessible to computational analysis
  - Readable by scientists
- **Establish ongoing efforts to curate (maintain, refine, embellish) these knowledge bases**

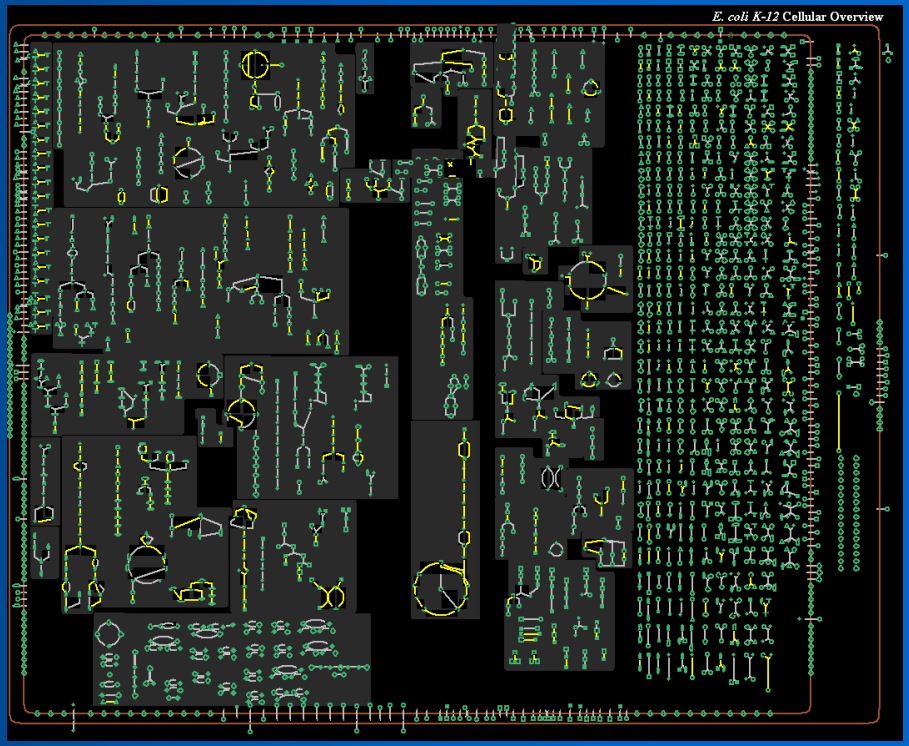
The Metabolic Network

- Chemical reactions interconvert chemical compounds
- An enzyme is a protein that accelerates a chemical reaction
- An enzyme is encoded by one or more genes in the genome
- A pathway is a linked set of reactions



The Metabolic Network

- **A large network of interconnected biochemical reactions**
- ***E. coli* metabolic network**
  - 220 transporters
  - 915 reactions (edges) involving 785 substrates (nodes)
  - 854 enzymes
- **The metabolic map of an organism is determined by the environments of that organism**

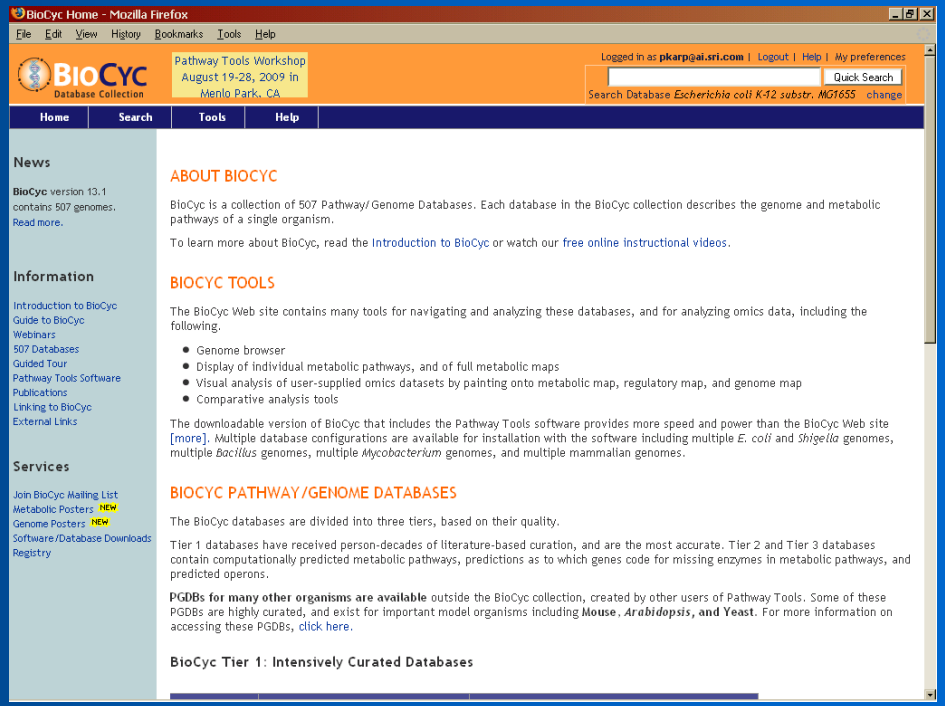


Pathway Tools Capabilities

- **Create and maintain an organism database integrating genome, pathway, regulatory information**
  - Computational inference tools
  - Interactive editing tools
- **Query and visualize that database**
- **Use the database to interpret omics data**
- **Metabolic network analysis tools**
- **Comparative analysis tools**
- **Export the metabolic network to SBML**
  - Speed creation of flux-balance models by order of magnitude

BioCyc Collection of 507 Pathway/Genome Databases

- **Pathway/Genome Database (PGDB) – combines information about**
  - Pathways, reactions, substrates
  - Enzymes, transporters
  - Genes, replicons
  - Transcription factors/sites, promoters, operons
- **Tier 1: Literature-Derived PGDBs**
  - MetaCyc
  - EcoCyc -- *Escherichia coli* K-12
- **Tier 2: Computationally-derived DBs, Some Curation -- 24 PGDBs**
  - HumanCyc
  - Mycobacterium tuberculosis
- **Tier 3: Computationally-derived DBs, No Curation -- 481 DBs**



Note

- **No formal connection to the Cyc project**

BioCyc Web Site Usage

- **BioCyc Web site receives 350,000 page views/month**

- ***E. coli* DB receives 150,000 page views/month**
- **MetaCyc DB receives 40,000 page views/month**

- **56,000 unique visitors in 2009**
- **20,000 visitors had at least 50 page views**

Alternate Availability of BioCyc

- **Data files downloadable for computational analysis**
- **Downloadable software/database bundle**
- **Third parties can deposit their PGDBs into PGDB registry for peer-to-peer sharing**

Curation

- **Computational inference quickly paints an approximate portrait of the organism**
- **For high value organisms with large scientific communities, paid curators are a worthwhile investment**
- **For other organisms, computational inference suffices**

MetaCyc: Metabolic Encyclopedia

- Describe a representative sample of every experimentally determined metabolic pathway
- Describe properties of metabolic enzymes

- Literature-based DB with extensive references and commentary
- Pathways, reactions, enzymes, substrates

- Jointly developed by
  - P. Karp, R. Caspi, C. Fulcher, SRI International
  - L. Mueller, A. Pujar, Boyce Thompson Institute
  - S. Rhee, P. Zhang, Carnegie Institution

Nucleic Acids Research **2010**

Applications of MetaCyc

- **Reference source on metabolic pathways**
- **Metabolic engineering**
  - Find enzymes with desired activities, regulatory properties
  - Determine cofactor requirements
- **Predict pathways from genomes**
- **Systematic studies of metabolism**
- **Computer-aided education**

MetaCyc Data -- Version 13.6

Pathways	1,436
Reactions	8,200
Enzymes	6,060
Small Molecules	8,400
Organisms	1,800
Citations	21,700

Taxonomic Distribution of MetaCyc Pathways – version 13.1

Bacteria	883
Green Plants	607
Fungi	199
Mammals	159
Archaea	112

MetaCyc Curation

- **DB updates by 5 staff curators**
  - Information gathered from biomedical literature
  - Emphasis on microbial and plant pathways
  - More prevalent pathways given higher priority
- **Review-level database**
- **Four releases per year**

- **Quality assurance of data and software:**
  - Evaluate database consistency constraints
  - Perform element balancing of reactions
  - Run other checking programs
  - Display every DB object

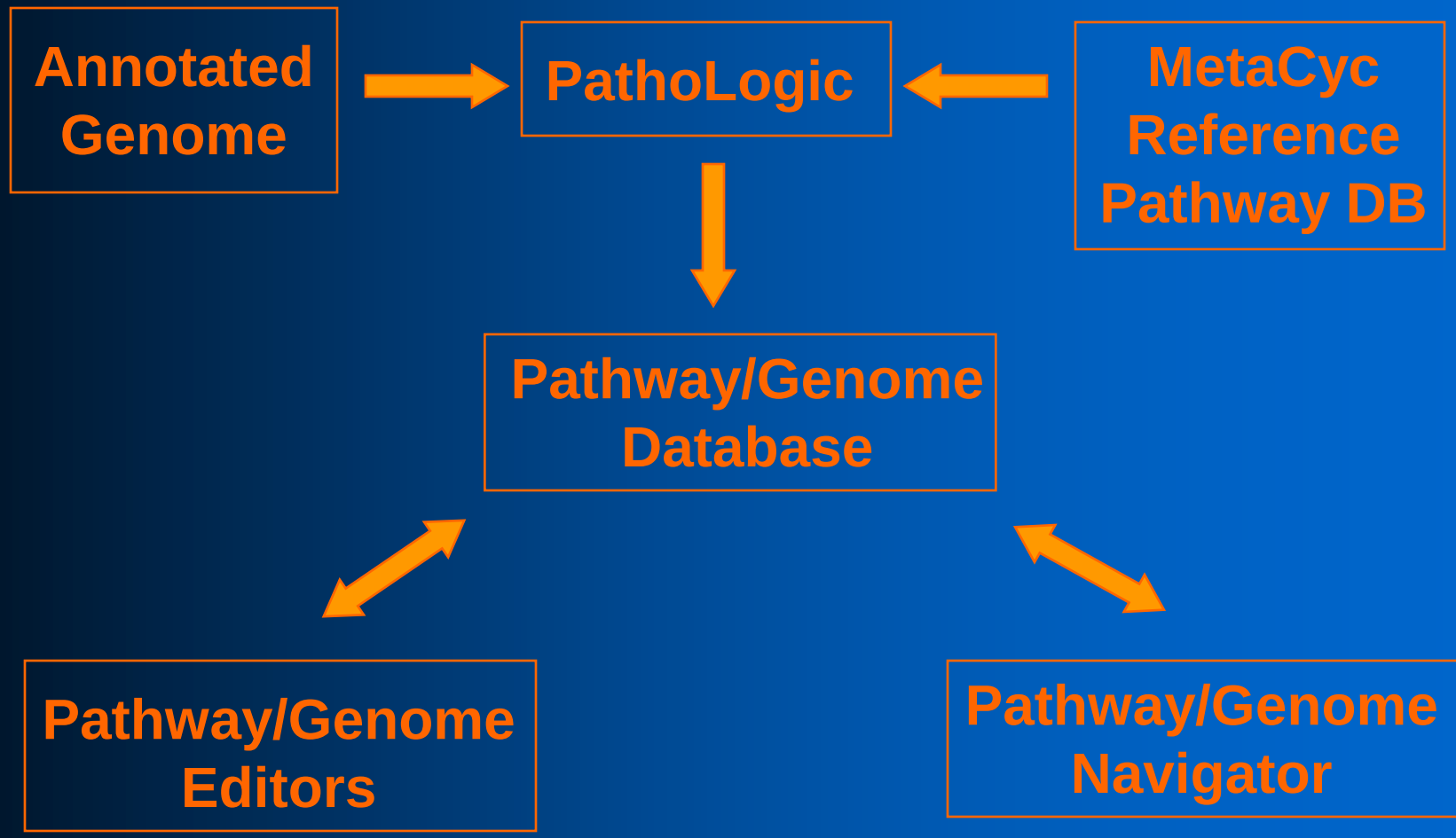
What is Curation?

- Ongoing updating and refinement of a PGDB
- Correcting false-positive and false-negative predictions
- Incorporating information from experimental literature
- Authoring of comments and citations
- Updating database fields
- Gene positions, names, synonyms
- Protein functions, activators, inhibitors
- Addition of new pathways, modification of existing pathways
- Defining TF binding sites, promoters, regulation of transcription initiation and other processes

Community-Based Curation?

- **Several significant experiments: Largely has not worked in the life sciences**
- **The people you most want to enter information are too busy**
- **Scientific culture does not reward people sufficiently for free contributions**
- **Biology PhD required to understand the knowledge**
- **Significant training required for**
  - Consistency of style
  - Ability to properly structure complex data

Pathway Tools Software Overview



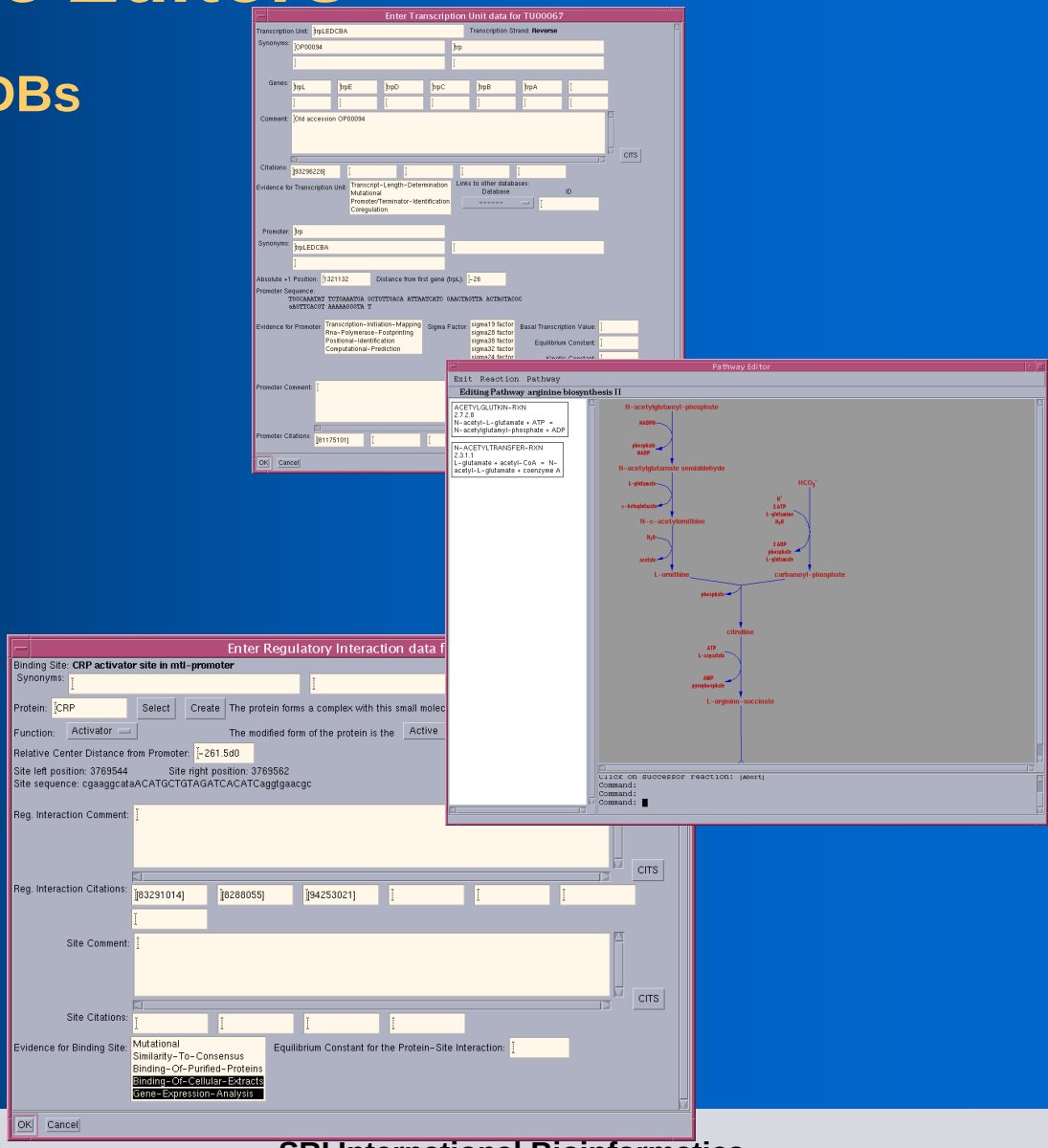
Pathway Tools Software: PathoLogic

- **Computational creation of new Pathway/Genome Databases**
- **Transforms genome into Pathway Tools schema and layers inferred information above the genome**
- **Predicts operons**
- **Predicts metabolic network**
- **Predicts which genes code for missing enzymes in metabolic pathways**
- **Infers transport reactions from transporter names**

Karp et al, *Briefings in Bioinformatics* 2009

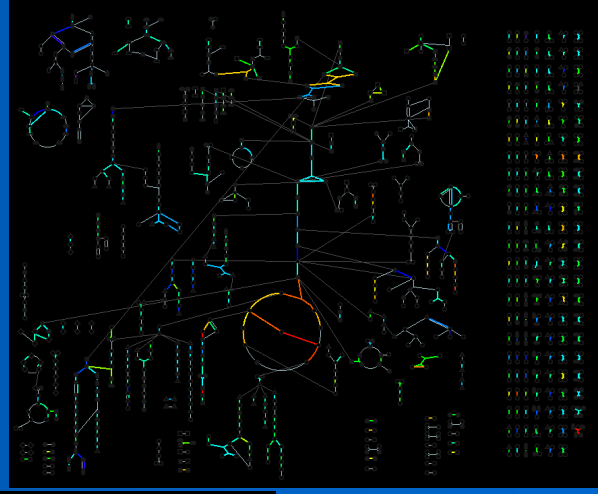
Pathway Tools Software: Pathway/Genome Editors

- Interactively update PGDBs with graphical editors
- Support geographically distributed teams of curators with object database system
- Gene editor
- Protein editor
- Reaction editor
- Compound editor
- Pathway editor
- Operon editor
- Publication editor

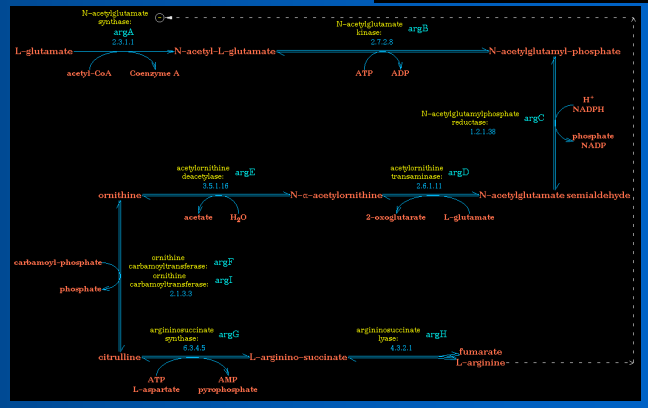


BioCyc Database Collection

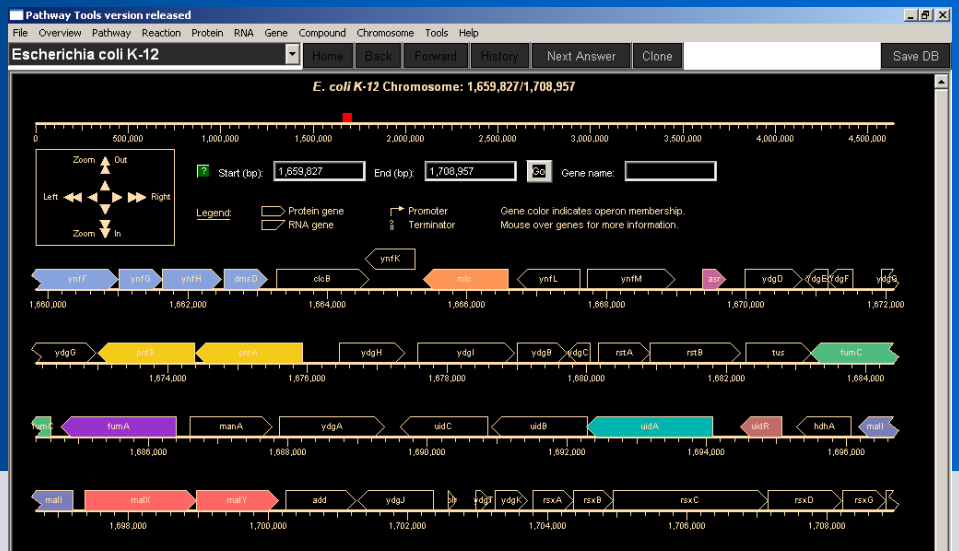
Pathway Tools Software: Pathway/Genome Navigator



- **Querying and visualization of:**
  - Pathways
  - Reactions
  - Metabolites
  - Genes/Proteins/RNA
  - Regulatory interactions
  - Chromosomes



- **Two modes of operation:**
  - Web mode
  - Desktop mode
  - Most functionality shared, but each has unique functionality



BioCyc Database Collection

Pathway Tools Implementation Details

- **Platforms:**
  - Macintosh, PC/Linux, and PC/Windows platforms
- **Same binary can run as desktop app or Web server**
- **Production-quality software**
  - Version control
  - Two regular releases per year
  - Extensive quality assurance
  - Extensive documentation
  - Auto-patch
  - Automatic DB-upgrade
- **480,000 lines of Lisp code**

Why Do We Code in Common Lisp?

- **Gatt studied Lisp and Java implementation of 16 programs by 14 programmers (Intelligence 11:21 2000)**
  - The average Lisp program ran 33 times faster than the average Java program
  - The average Lisp program was written 5 times faster than the average Java program
- **Roberts compared Java and Lisp implementations of a Domain Name Server (DNS) resolver**
  - http://www.findinglisp.com/papers/case_study_java_lisp_dns.html
  - The Lisp version had ½ as many lines as code

Peter Norvig's Solution

- **"I wrote my version in Lisp. It took me about 2 hours (compared to a range of 2-8.5 hours for the other Lisp programmers in the study, 3-25 for C/C++ and 4-63 for Java) and I ended up with 45 non-comment non-blank lines (compared with a range of 51-182 for Lisp, and 107-614 for the other languages). (That means that some Java programmer was spending 13 lines and 84 minutes to provide the functionality of each line of my Lisp program.)"**

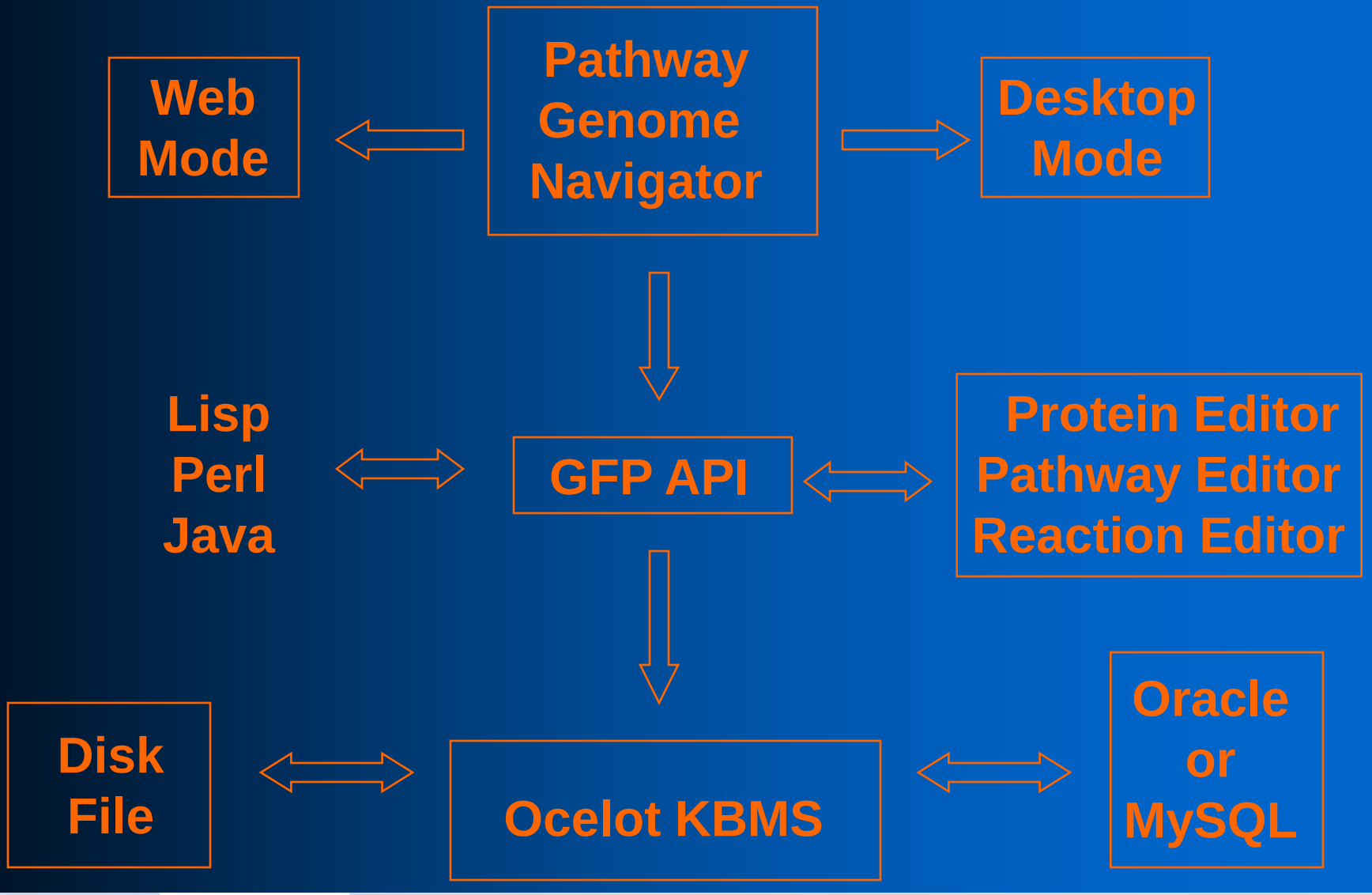
- **http://www.norvig.com/java-lisp.html**

Pathway Tools Schema / Ontology

- **1064 classes**
  - Datatype classes such as:
    - Pathways, Reactions, Compounds, Macromolecules, Proteins, Replicons, DNA-Segments (Genes, Operons, Promoters)
  - Taxonomies for Pathways, Reactions, Compounds
  - Cell Component Ontology
  - Evidence Ontology

- **277 attributes and relationships**
  - Meta-data: Creator, Creation-Date
  - Comment, Citations, Common-Name, Synonyms
  - Attributes: Molecular-Weight, DNA-Footprint-Size
  - Relationships: Catalyzes, Component-Of, Product

Pathway Tools Architecture



The Ocelot Data Model

- **Frame data model reduces schema complexity**
- **Frames are of two types: classes, instances**

- **Frames have slots that define their properties, attributes, relationships**
- **A slot has one or more values**
- **Each value can be any Lisp datatype**

- **Slotunits define metadata about slots:**
  - Domain, range, inverse
  - Collection type, number of values, value constraints

Inference Capabilities

- **Inheritance of defaults**
- **Slot values computed via attached procedures**
- **Maintenance of inverse relationships**
- **No classifier**
- **Constraint system**
  - Deferred evaluation
  - Tolerant of nonconformant data

Ocelot Storage System Architecture

- **Persistent storage via disk files or Oracle or MySQL**
  - Concurrent development: Oracle or MySQL
  - Single-user development: disk files
- **Oracle/MySQL DBMS storage**
  - DBMS is submerged within Ocelot, invisible to users
  - Relational schema is domain independent, supports multiple KBs simultaneously
  - Frames are cached in memory
  - Frames transferred from DBMS to Ocelot
    - On demand
    - By background prefetcher
    - Persistent disk cache to speed performance via Internet
- **Transaction logging facility**

Transaction Logging

- **Relational DBMS stores**
  - The latest version of each Ocelot frame
  - A log of all GFP operations applied to KB
- **Transaction log enables:**
  - Reconstruction of earlier versions of KB
  - View history of changes to an object
  - Update replicates of a KB
  - Detection of update conflicts during concurrency control
  - Undo of updates

Optimistic Concurrency Control

- **Locking approach**
  - Updates to one object can require locking all connected objects, and can cascade to thousands of objects
  - Locking complicates application logic
- **No locking**

- **User performs updates in local workspace**

- **When user commits changes, storage system compares user changes against all other committed changes**

Ocelot Knowledge Server Schema Evolution

- **FRSs store and process class and instance information similarly**
- **Application can query schema information as easily as it can query instances**

- **Schema is stored within the DB**
- **Schema is self documenting**
- **Schema evolution facilitated by**
  - Easy addition/removal of slots, or alteration of slot datatypes
  - Flexible data formats that do not require dumping/reloading of data

Ocelot Results

- Ocelot in use with MetaCyc, EcoCyc, and BioCyc projects for ~13 years
- Manages 500 BioCyc PGDBs of ~10-20K frames each
- MetaCyc PGDB contains 88K frames
- HumanCyc PGDB contains 117K frames

- Supports collaborative development of EcoCyc and MetaCyc by five groups in North America, Australia, and Mexico

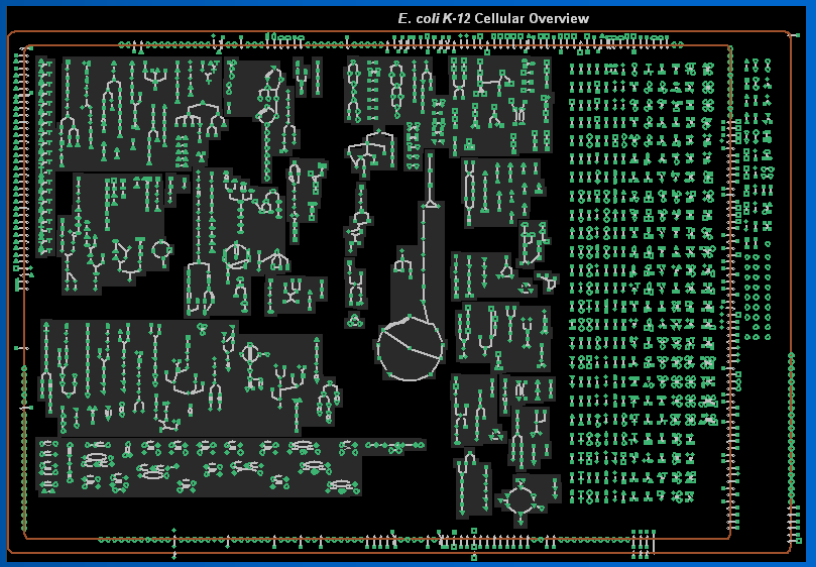
New Allegro Cache Based Storage Back-End

- A range of performance experiments in hand
- Results indicate it will scale to thousands of PGDBs

Tools

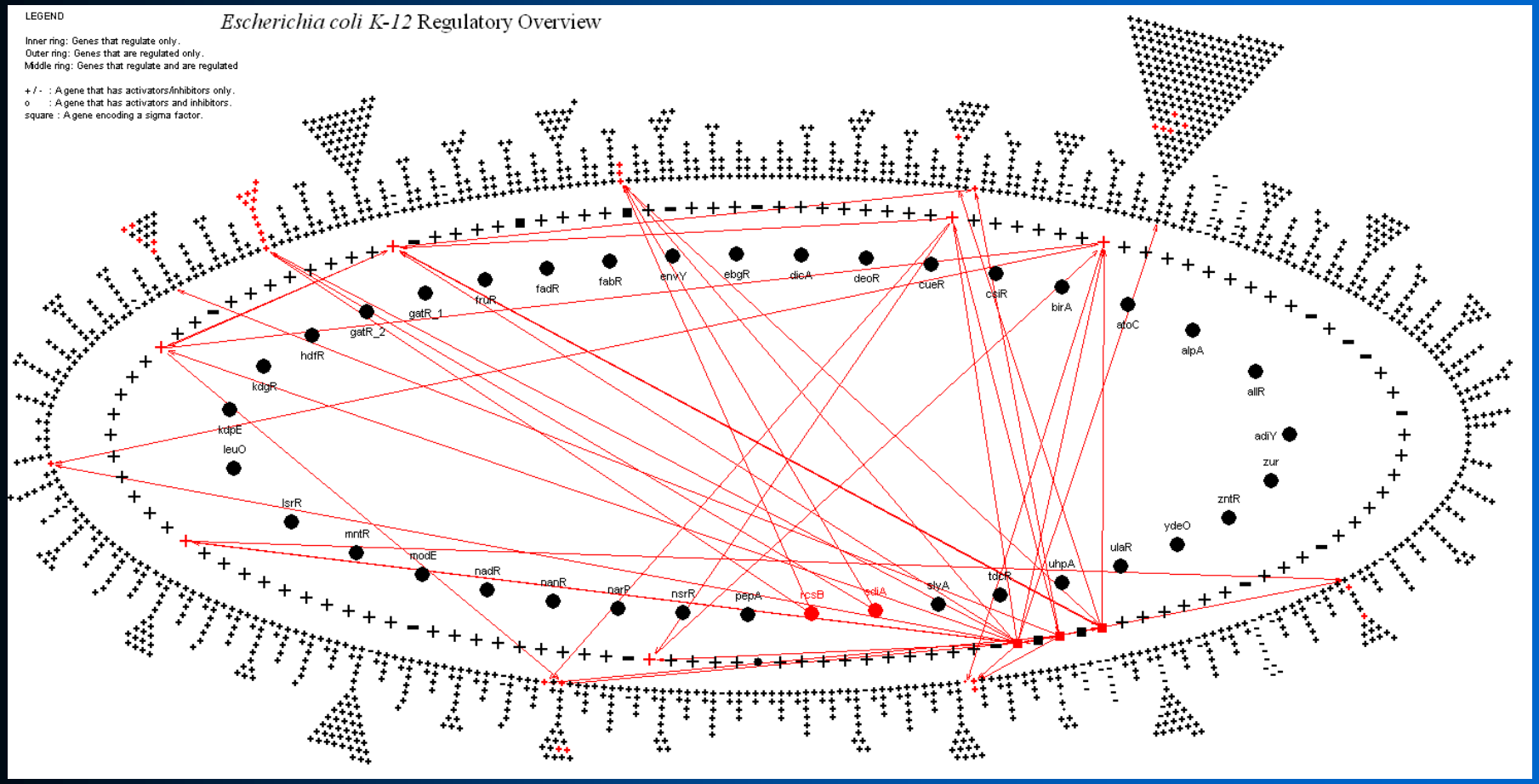
Pathway Tools Overviews and Omics Viewers

- **Genome-scale visualizations of cellular networks**
- **Harness human visual system to interpret patterns in biological contexts**
- **Designed to avoid the hairball effect**
- **Generated automatically from PGDB**
- **Magnify, interrogate**
- **Omics viewers paint omics data onto overview diagrams**
  - Different perspectives on same dataset
  - Use animation for multiple time points or conditions
  - Paint any data that associates numbers with genes, proteins, reactions, or metabolites



Regulatory Overview and Omics Viewer

- Show regulatory relationships among gene groups



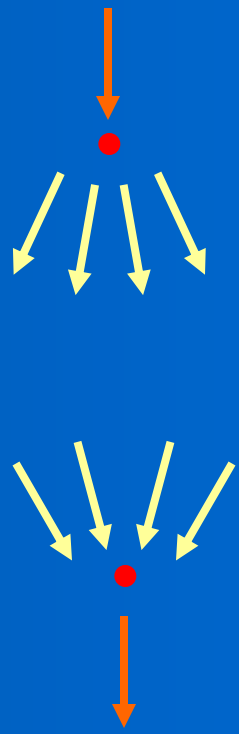
BioCyc Database Collection

Genome Overview

BioCyc Database Collection

Infer Anti-Microbial Drug Targets

- **Infer drug targets as genes coding for enzymes that encode chokepoint reactions**
- **Two types of chokepoint reactions:**

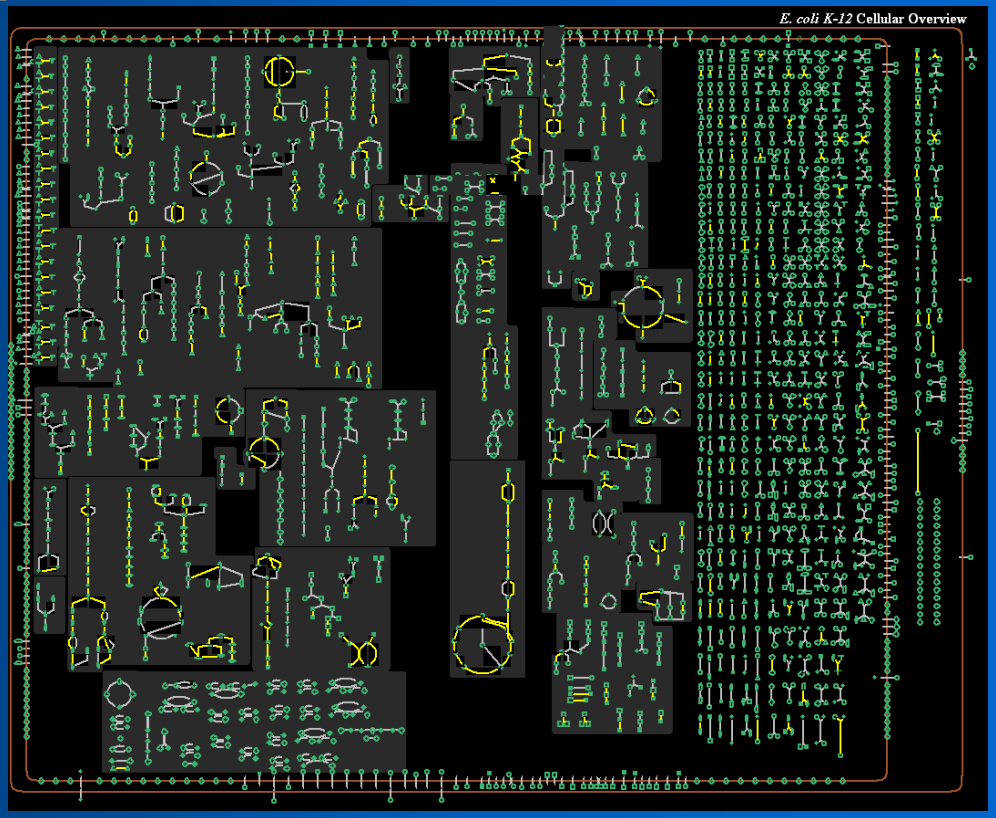


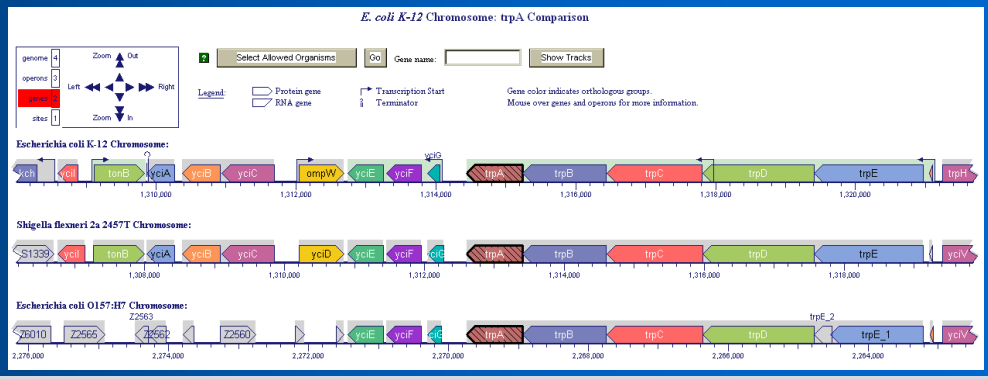
- **Chokepoint analysis of *Plasmodium falciparum*:**
  - 216/303 reactions are chokepoints (73%)
  - All 3 clinically proven anti-malarial drugs target chokepoints
  - 21/24 biologically validated drug targets are chokepoints
  - 11.2% of chokepoints are drug targets
  - 3.4% of non-chokepoints are drug targets
  - => Chokepoints are significantly enriched for drug targets

***Genome Research* 14:917 2004**

Comparative Analysis

- Via Cellular Overview
- Comparative genome browser
- Comparative pathway table
- Comparative analysis reports
  - Compare reaction complements
  - Compare pathway complements
  - Compare transporter complements





BioCyc Database Collection

SRI International Bioinformatics

Authoring Precise Database Queries

- **How to enable scientists to perform complex SQL-like queries?**
- **Structured Advanced Query Form**
  - Interactive form that guides user in constructing query
  - Javascript implementation
  - Query is pseudo-readable
  - Selectors present user with valid options at every point
    - Such as based on datatype of a given slot
    - Protects user from building malformed queries
  - Schema driven
    - Entire PGDB schema sent to Web browser via XML
  - User need not learn a query language

Structured Advanced Query Form

1. **Select database to query**
2. **Select class to query**
3. **Define one or more conditions**
   - Key advance: Conditions can navigate to related objects
4. **Select attributes for query output**
5. **Select output data format**

SAQP User Evaluation

- **8 biologists asked to answer 10 questions using SAQP during session at SRI**
  - "Find all E. coli pathways having more than 2 reactions where no reactions contain acetaldehyde"
- **One group had no training; one group watched 10 minute video tutorial on SAQP**
- **Results:**
  - 25% of subjects answered 75% of questions
  - 38% of subject answered 50% of questions
  - Highest scores from two participants who watched video and one participant with Perl programming experience
  - Congruence between terminology in questions vs terminology in schema a key factor, as is understanding of schema itself

Summary

- **BioCyc collection of Pathway/Genome Databases**
  - Most PGDBs created computationally
  - Professional curation staff added information from biomedical literature
- **Large user community accesses browsable Web site**
- **Large ontology provides high fidelity representation and compatibility of dataqbases**

Summary

- **KB development supported by Ocelot KBMS**
  - Persistent KB storage in Oracle and MySQL
  - Transaction log supports optimistic concurrency control
  - Disk cache allows collaboration across continents
- **Common Lisp development has allowed a small team to build and maintain a large code base**
- **Interactive query builder allows scientists to construct precise queries without SQL**
- **Extensive suite of visualization and analysis applications**

Acknowledgements

- **SRI**
  - Suzanne Paley, Ron Caspi, Ingrid Keseler, Carol Fulcher, Markus Krummenacker, Alex Shearer, Tomer Altman, Fred Gilham, Pallavi Kaipa
- **EcoCyc Collaborators**
  - Julio Collado-Vides, Robert Gunsalus, Ian Paulsen
- **MetaCyc Collaborators**
  - Sue Rhee, Peifen Zhang, Kate Dreher
  - Lukas Mueller, Anuradha Pujar
- **Funding sources:**
  - NIH National Center for Research Resources
  - NIH National Institute of General Medical Sciences
  - NIH National Human Genome Research Institute

BioCyc.org

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