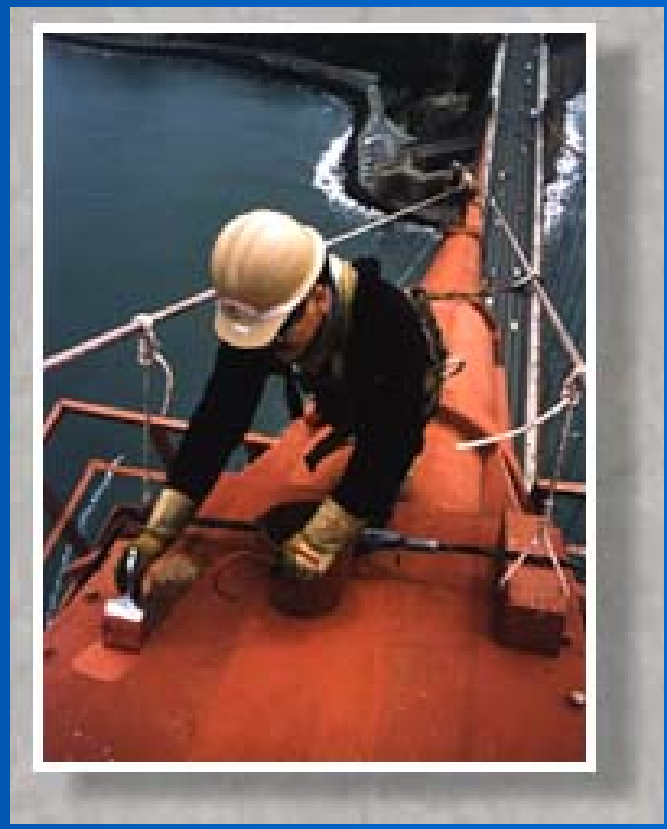
The EcoCyc *Curation Process*

Ingrid M. Keseler

SRI International

HOW OFTEN IS THE GOLDEN GATE BRIDGE PAINTED?

Many misconceptions exist about how often the Bridge is painted. Some say once every seven years, others say from end-to-end each year. Actually, the Bridge was painted when it was originally built. For the next 27 years, only touch up was required. By 1965, advancing corrosion sparked a program to remove the original paint and replace it with an inorganic zinc silicate primer and acrylic emulsion topcoat. The program was completed in 1995. The Bridge will continue to require routine touch up painting on an on-going basis.



http://goldengatebridge.org/research/facts.php

EcoCyc – a History of Change

- **EcoCyc began as a database of *E. coli* metabolic pathways**
- **Goals for EcoCyc curation have changed**
  - Annotate all *E. coli* gene products
- **Tools for EcoCyc curation have changed**
  - Changes in Pathway Tools allow entry of additional data types
- **Curators have changed**
  - Approximately 10 former contributors
- **Even the *E. coli* genome has changed**
  - Updated and corrected *E. coli* sequence June `04
  - Updated and corrected *E. coli* genes and annotations Jan. `06

Who?
Why?
What?
How?

EcoCyc Project – EcoCyc.org

- ***E. coli* Encyclopedia**
  - Model-Organism Database for *E. coli*
  - Tracks evolving annotation of the *E. coli* genome

- **Collaborative development via Internet**
  - John Ingraham (UC Davis)
  - Paulsen (TIGR) – Transport, flagella, DNA repair, …
  - Collado (UNAM) -- Regulation of gene expression
  - Keseler, Shearer (SRI) -- Metabolic pathways, cell division, proteases, RNases, replication, ribosome, …
  - Karp (SRI) -- Bioinformatics

Why?

Who is using EcoCyc?

- **Experimentalists**
  - *E. coli* experimentalists
  - Experimentalists working with other microbes
  - Analysis of expression data
- **Computational biologists**
  - Biological research using computational methods
  - Genome annotation
  - Study connectivity of *E. coli* metabolic network
  - Study organization of *E. coli* metabolic enzymes into structural protein families
  - Study phylogenetic extent of metabolic pathways and enzymes in all domains of life
- **Bioinformaticists**
  - Training and validation of new bioinformatics algorithms – predict operons, promoters, protein functional linkages, protein-protein interactions
- **Metabolic engineers**
  - "Design of organisms for the production of organic acids, amino acids, ethanol, hydrogen, and solvents "
- **Educators**

What?

Highlights of the EcoCyc Data

- **Extensive commentary and literature citations**
- **Ongoing extensive literature curation effort to update all of the following types of data**
- **Genome**
  - Gene names, synonyms, positions, sequence, interrupted?, paralogs
- **Proteome and stable RNAs**
  - Product names, synonyms, subunit organization, citations, comments, pI, molecular weight, sequence
  - Enzyme substrate specificity, activators, inhibitors, cofactors
  - Transporter substrates
  - Transcription factor binding sites, interactions

Highlights of the EcoCyc Data

- **Metabolic pathways**
  - Names, Comments, citations, pathway links, superpathways, reaction topology
- **Metabolic and transport reactions**
  - Substrates, EC number, spontaneous?
- **Small molecules**
  - Names, chemical structures
- **Genetic network**
  - Operons, promoters, transcription-factor binding sites, regulatory reactions
- **Taxonomies (Ontologies)**
  - Genes (MultiFun, GO), Pathways, Compounds, Reactions
- **Database links (Swiss-Prot, CGSC, PIR, PDB, Swiss-Model, ModBase, RefSeq, EcoGene)**

How?

Literature-Based Curation

- **EcoCyc incorporates information from the published literature**
  - Functions may be experimentally or computationally assigned; evidence codes are used to distinguish these
- **What EcoCyc is not:**
  - EcoCyc does not currently perform its own function predictions for gene products
  - EcoCyc is not a repository for large-scale gene expression, proteomic or other high-throughput data

Examples

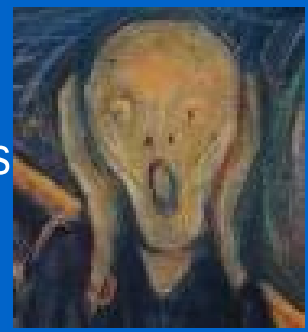
Literature Searches

- **Searching PubMed for "coli" gives many spurious results:**
  - 254,447 references as of June 13, 2006
  - Only ~10-20% refer to *E. coli* K-12 as the experimental organism – there are many publications on pathogenic strains and using *E. coli* for protein expression

Literature Searches

Literature Searches

- **Searching PubMed for "coli" gives many spurious results:**
  - 254,345 references as of June 9, 2006
  - Only ~10-20% refer to *E. coli* K-12 as the experimental organism – there are many publications on pathogenic strains and using *E. coli* for protein expression
  - …and that still leaves at least 25,000 publications!!!
  - ~ 200 new *E. coli* papers per month
- **Searching by gene name, protein name, synonyms**
  - …but how to search for information on genes named *lasT*? *lasT* AND coli = 2134 publications!
- **Automatic PubMed searches at My NCBI**



Help with Literature Searches

- **Literature update summaries by an expert**
  - Edward Adelberg (Professor Emeritus at Yale) for *E. coli* – now done by Narinder Whitehead
- **Other databases**
  - EcoGene, UniProt, GenProtEC, EchoBASE, …
- **For subjects with vast literature resources: review articles, EcoSal**

Strategies for Curation

- **Curation by process or system**
  - Proteases, RNases, DNA polymerases, ribosome, transporters, pathways, …

  Advantages:
  - More efficient literature searches
  - Reviews often cover entire systems
  - Enables “Big Picture” view of the organism

Strategies for Curation

- **Curation by "available expert"**
  - Recruit experts who will assist in the curation effort by suggesting publications, reviewing the curator's work, or writing comments

  Advantages:

  - Pre-digested literature
  - Expert review ensures correctness (one hopes)
  - Small time commitment by expert

Strategies for Curation

- **Curation by novelty**
  - Review newly published literature for novel function assignments

  Advantages:
  - New literature will be quickly incorporated in the database
  - New function assignments will help annotation of other proteins

Strategies for Curation

- **Gene-by-gene curation**
  - Systematic literature searches using a list of genes (by chromosome position, by date they were last updated, etc.)

  Advantages:
  - Finding things that were previously missed
  - Keeping the list of “unstudied” gene products up to date

Coordination, Communication and Documentation

- **Coordination and regular communication between curators is essential, especially if some are located off-site**
- **Written documentation: Curator's Guide**
  - Updated as priorities and strategies change

How do you set up a new PGDB curation project?

- **Estimating the size of the project**
  - Size of the current literature
  - Rate of new publications
  - Scope of the project
- **Acquiring funding**
  - Panel discussion
- **Hiring curators**
  - Who is the right person for the job?
  - How much experience is needed/desirable?

Acknowledgements

- **Bioinformatics Research Group, SRI**
- **Collaborators:**
  - John Ingraham
  - Paulsen group
  - Collado group
- **Previous curators**
- **Funding**
  - NIH National Center for Research Resources