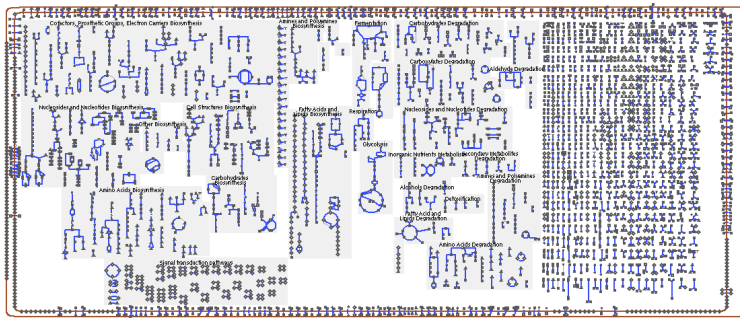
SRI International



Accelerating Science with BioCyc AND Computational Challenges from the Human Microbiome

Peter D. Karp
SRI International

ecocyc.org
biocyc.org
metacyc.org

BioCyc™ Database Collection

EcoCyc™

Overview

- Overview of our bioinformatics software and databases
- Modeling the human microbiome

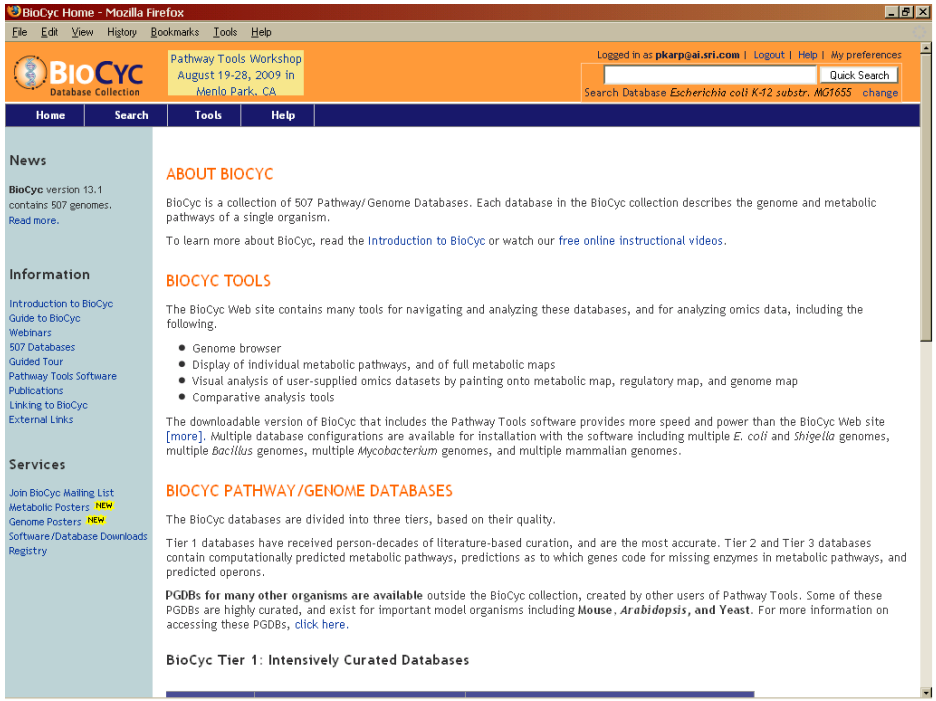
"Big Knowledge": BioCyc.org Collection of 7,600 Pathway/Genome Databases

- Pathway/Genome Database (PGDB) – combines information about
  - Pathways, reactions, substrates
  - Enzymes, transporters
  - Genes, replicons
  - Transcription factors/sites, promoters, operons

- Tier 1: Highly curated PGDBs
  - MetaCyc, HumanCyc, YeastCyc
  - EcoCyc -- *Escherichia coli* K-12
  - AraCyc – *Arabidopsis thaliana*

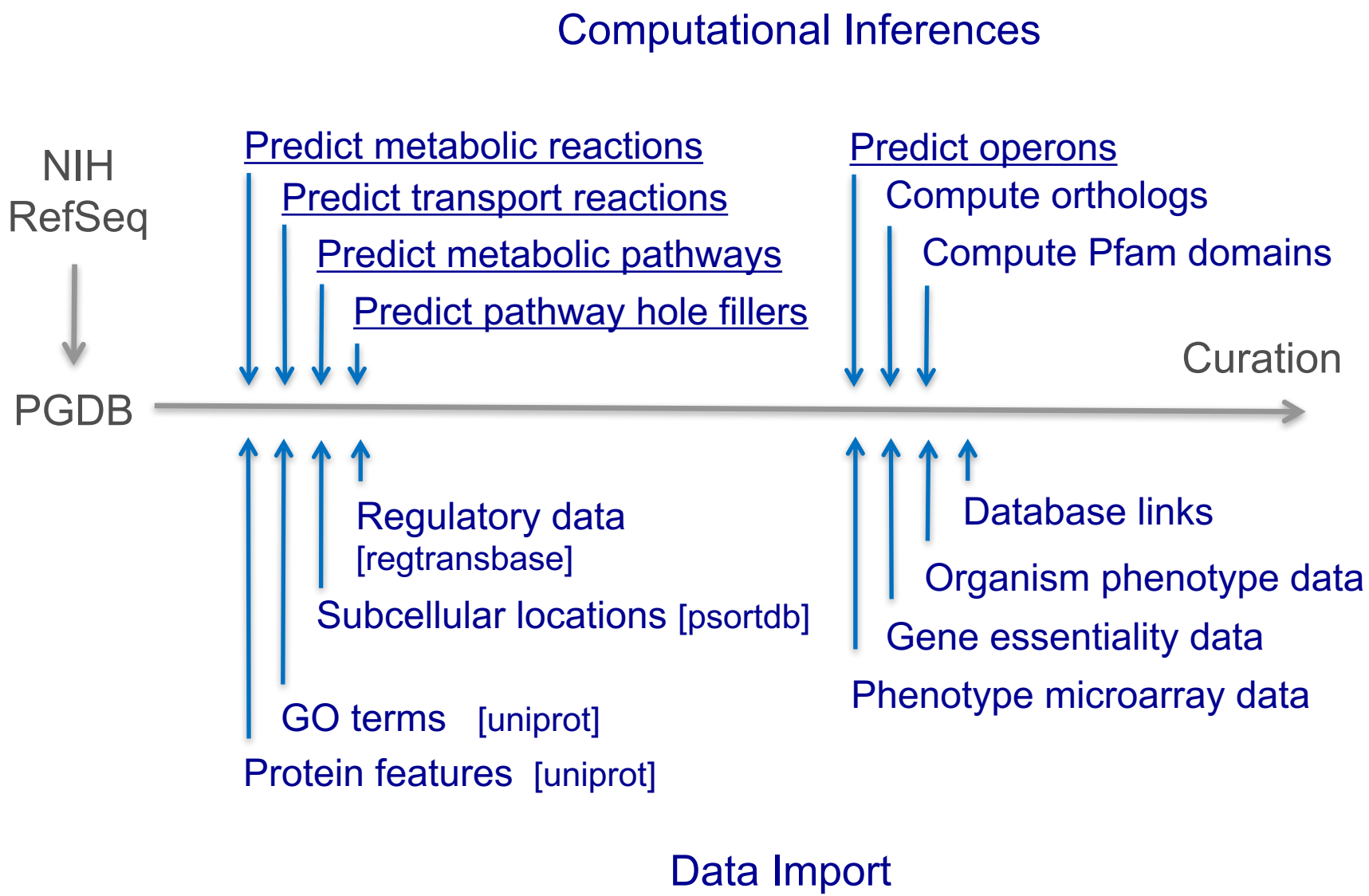
- Tier 2: Moderately curated -- 44 PGDBs
  - *Bacillus subtilis, Mycobacterium tuberculosis*

- Tier 3: Computationally-derived DBs, No Curation -- ~7,600 PGDBs

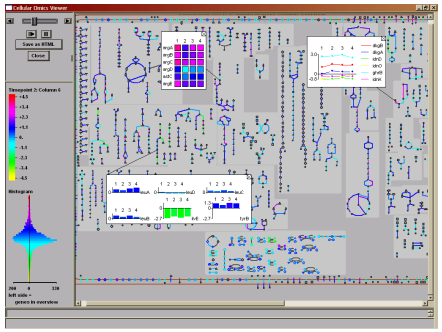


- BioCyc content derived from:
  - Other databases
  - Computational inferences
  - 60,000 curated publications
- ~1M page views/month

Creation of BioCyc Databases



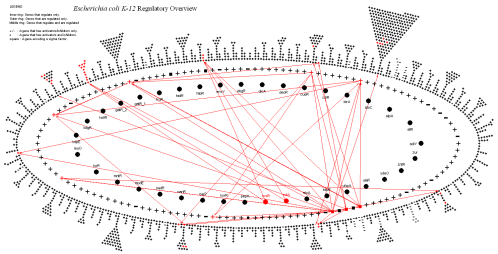
Pathway Tools



Annotated Genome

+

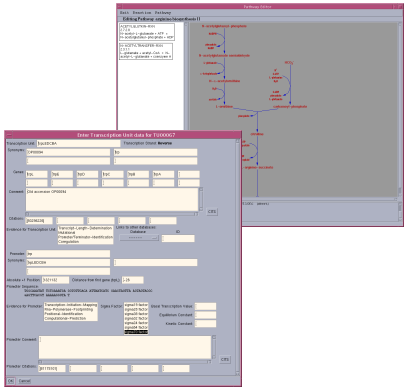
PathoLogic MetaCyc



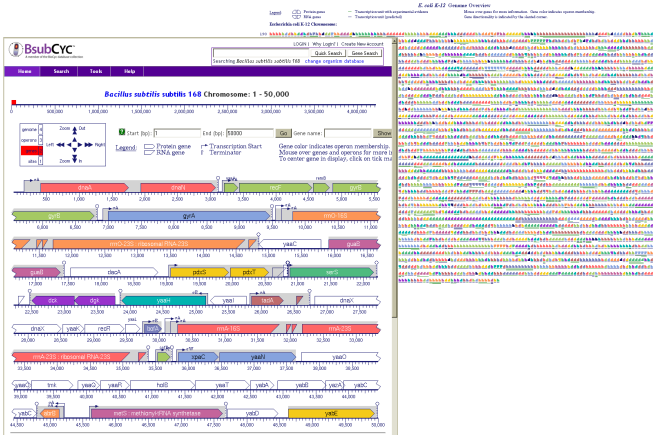
MetaFlux

Pathway/Genome Database

Pathway/Genome Navigator

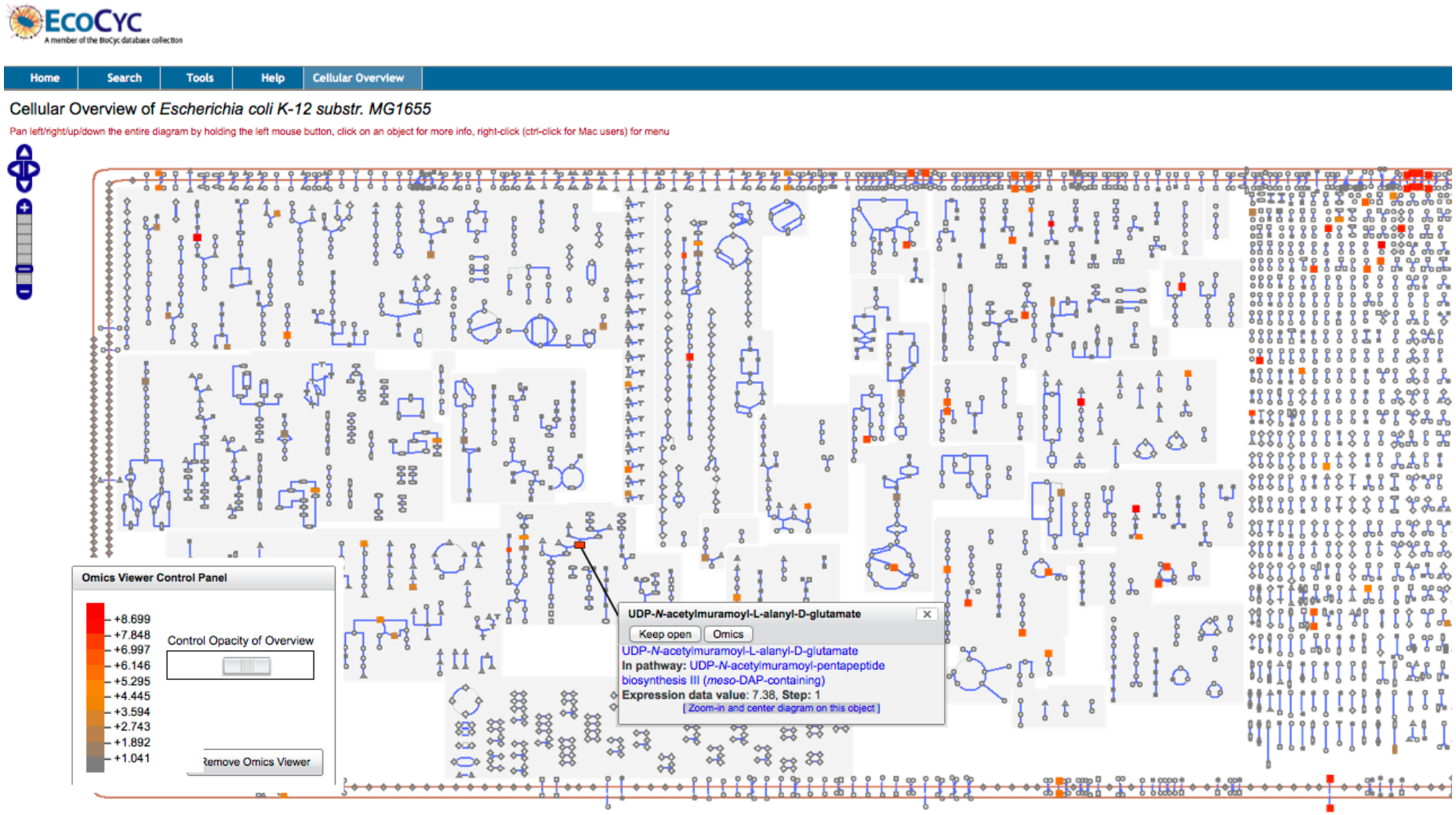


Pathway/Genome Editors



Licensed by 7,000+ Groups

Metabolomics Data Painted on Metabolic Map



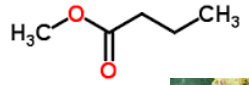
Human Gut Microbiome

- ~ 1,000 bacteria, archaea, fungi
- Constitutes a distinct human organ
- What mechanisms underlie operation of the gut microbiome?
- What interventions can bring the microbiome back to a healthy state?

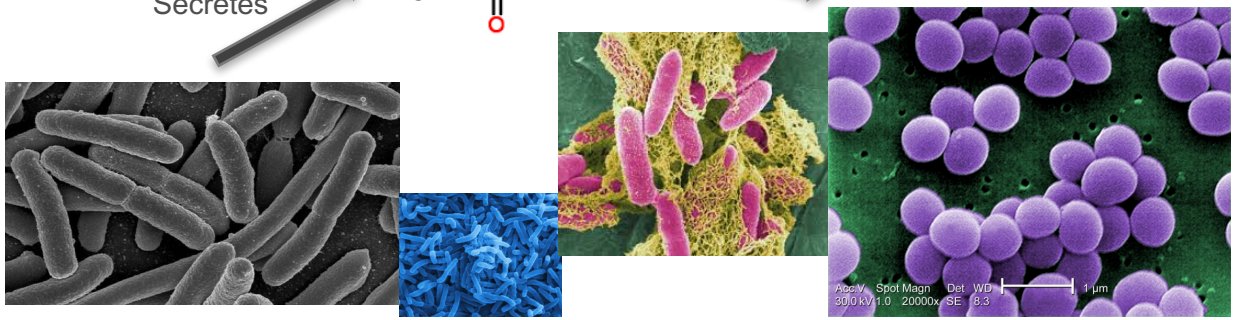
Microbiome Imbalances Linked to Diseases

Healthy State

Secretes



Uptakes



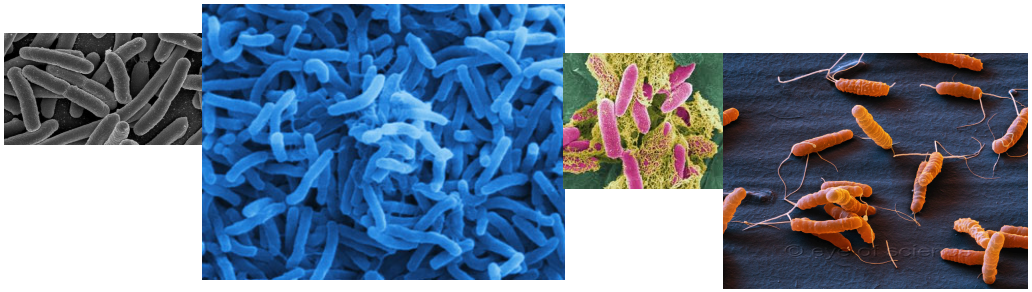
Diet

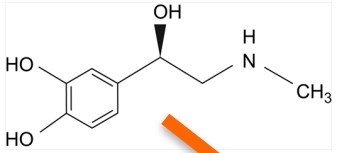
Probiotics

Antibiotics

Diet

Diseased State





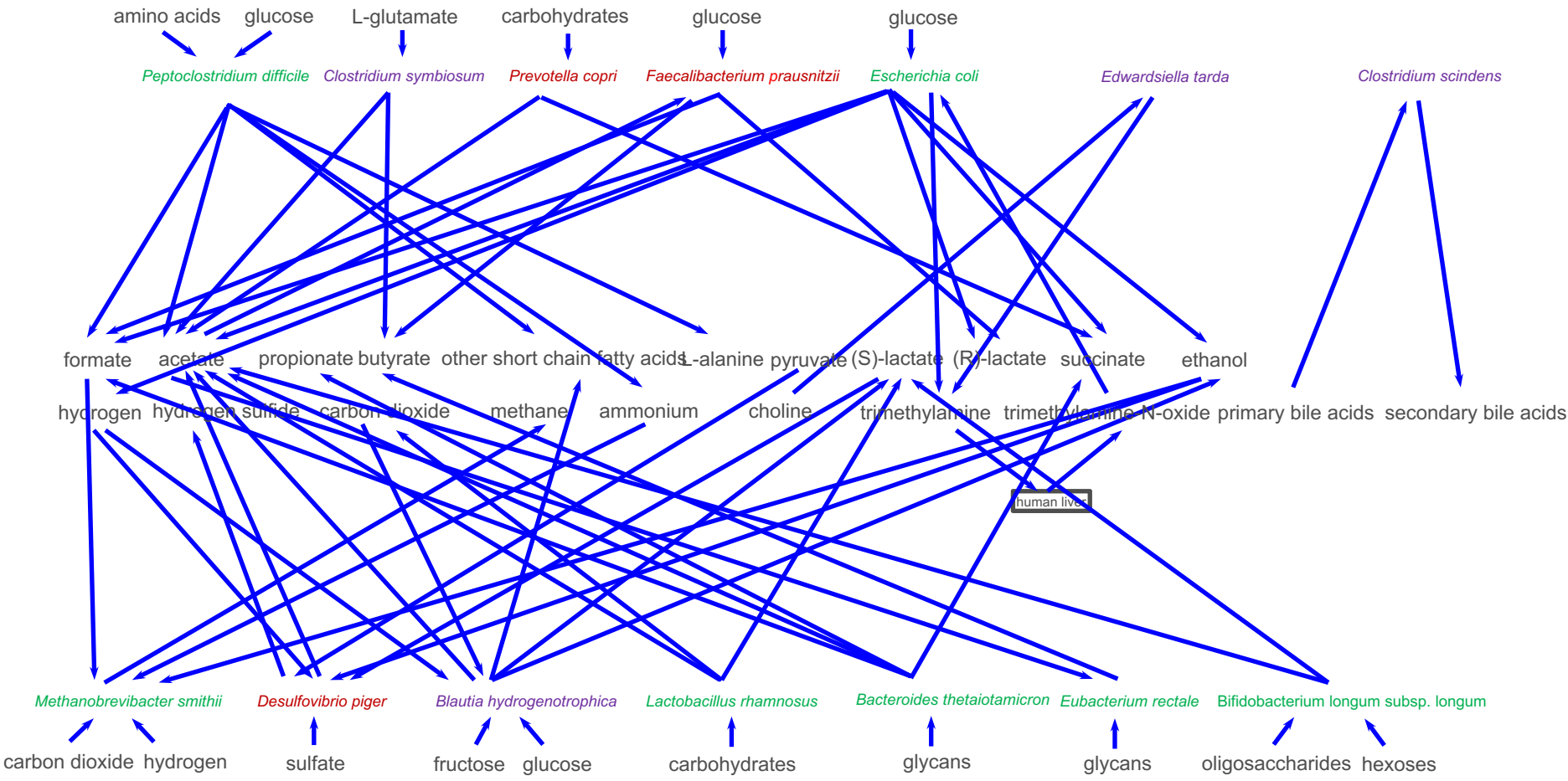
Brain

Immune System

The Microbiome Ecosystem

Bioinformatics Research Group, SRI International, Menlo Park CA USA

BioCyc.org



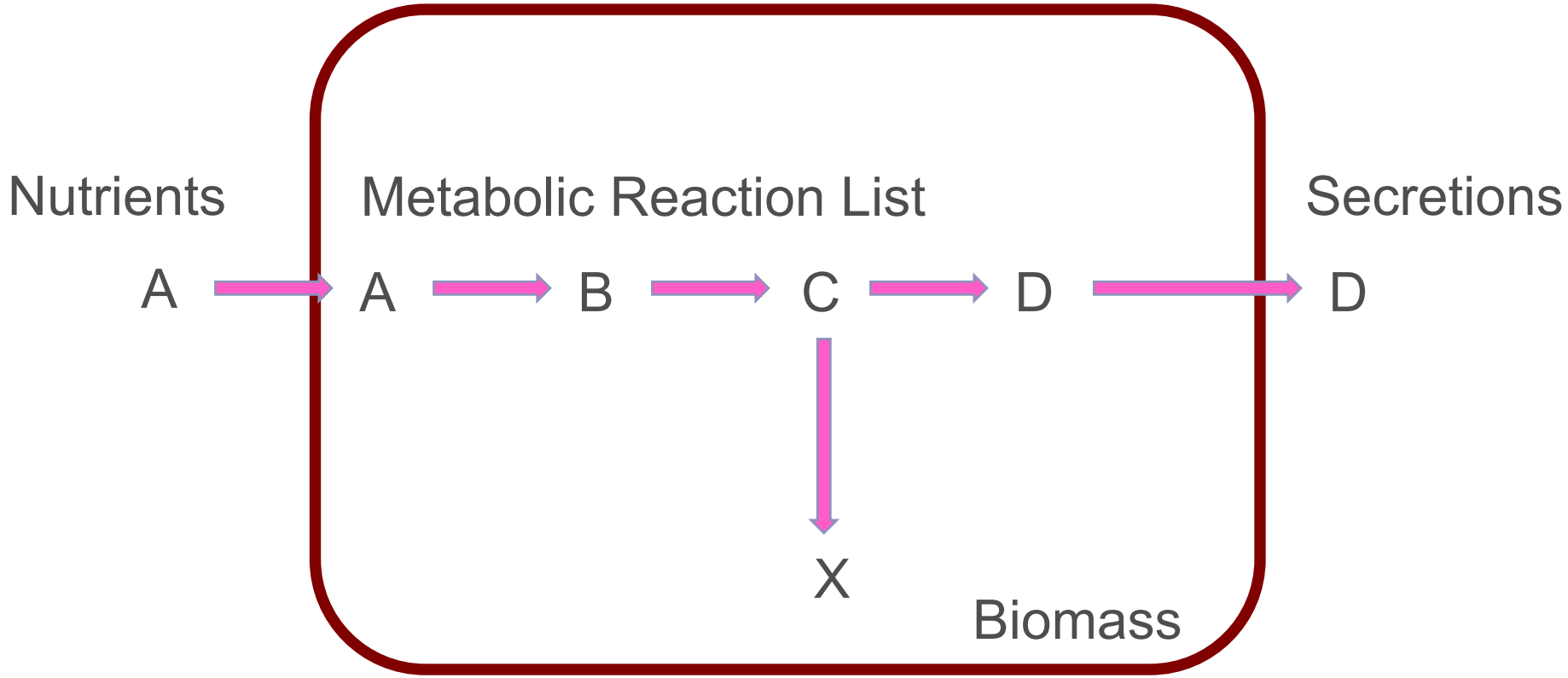
SRI genome scale metabolic models

partial models

potential models

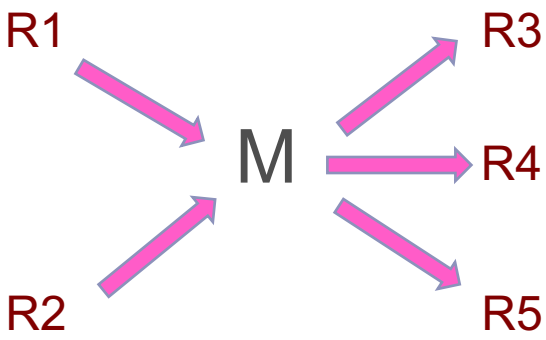
Metabolic Modeling

- Constraint-based quantitative models of metabolism
- *E. coli* model derived from EcoCyc database (*BMC Sys Biol* 2014 8:79):
  - 16 nutrients
  - 108 biomass metabolites
  - 2286 reactions



Flux Balance Analysis

- Define system of linear equations encoding fluxes on each metabolite M
  - R1 + R2 = R3 + R4 + R5



- Boundary reactions:
  - Exchange fluxes for nutrients and secretions
  - Biomass reaction L-arginine … + GTP … + … → biomass

- Submit to linear optimization package
  - Optimize biomass production or
  - Optimize ATP production or
  - Optimize production of desired end product

- Output: Assigned fluxes (rates) for every reaction

MetaFlux Modeling Tool: Modes of Operation

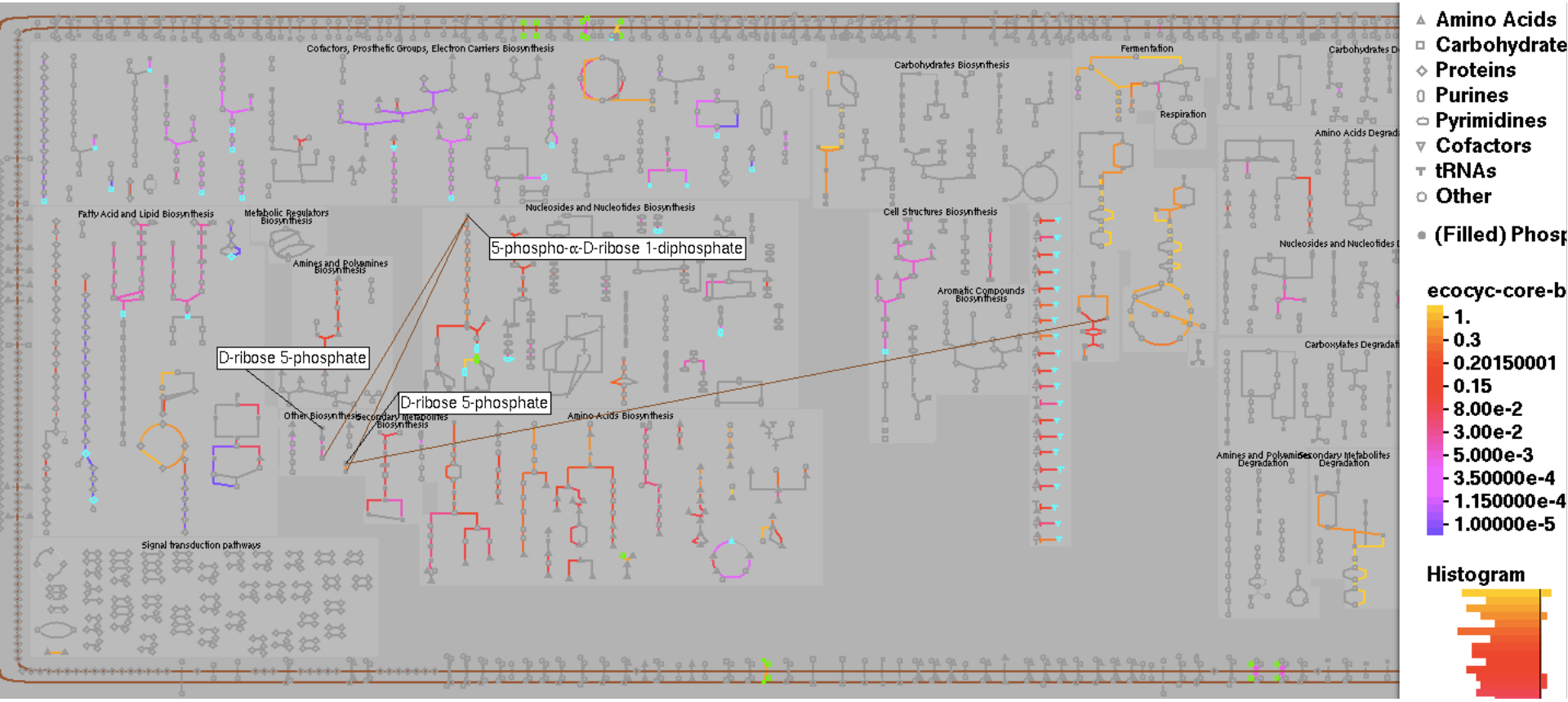
- **Solving mode**
  - Individual organisms, organism communities
  - Steady-state FBA, dynamic FBA
  - Single compartment, 2-D spatial grid with diffusion
  - Removal of flux loops
- **Knock-out mode** (single/double gene/reaction knock-outs)
- **Model development mode**
  - Inference of biomass reaction
  - Development mode (multiple gap filling)
    - Suggests reactions to add; identifies non-producible biomass metabolites

Karp *et al*, *Briefings in Bioinformatics*, 2015

E. coli Metabolic Model

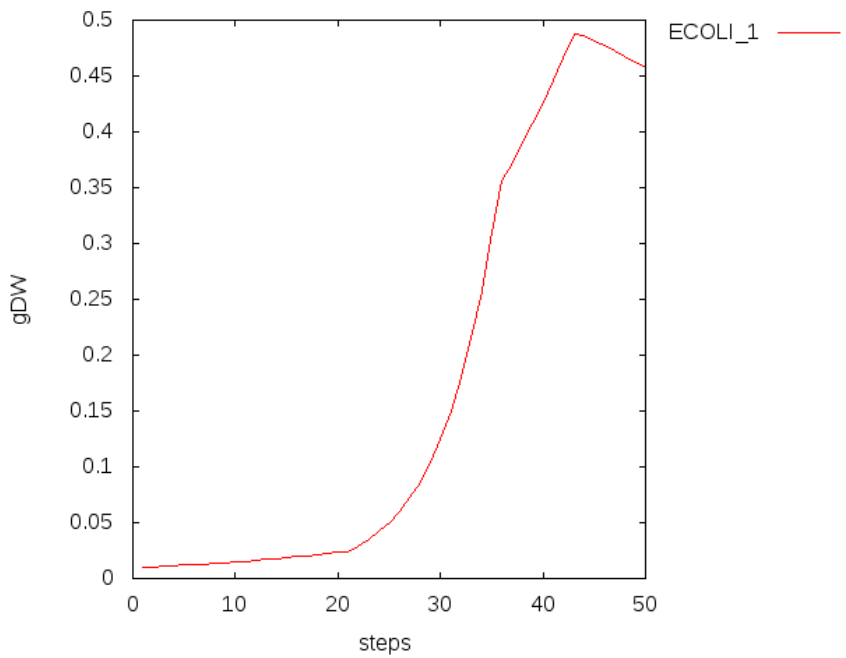
- Generated from EcoCyc, updated on each EcoCyc release
- Predicts phenotypes of *E. coli* knock-outs
  - 95.2% accuracy for 1445 genes
- Predicts growth/no-growth of *E. coli* on different nutrients
  - 80.7% accuracy for 431 chemically defined growth media
- *BMC Syst Biol.* 2014 Jun 30;8:79

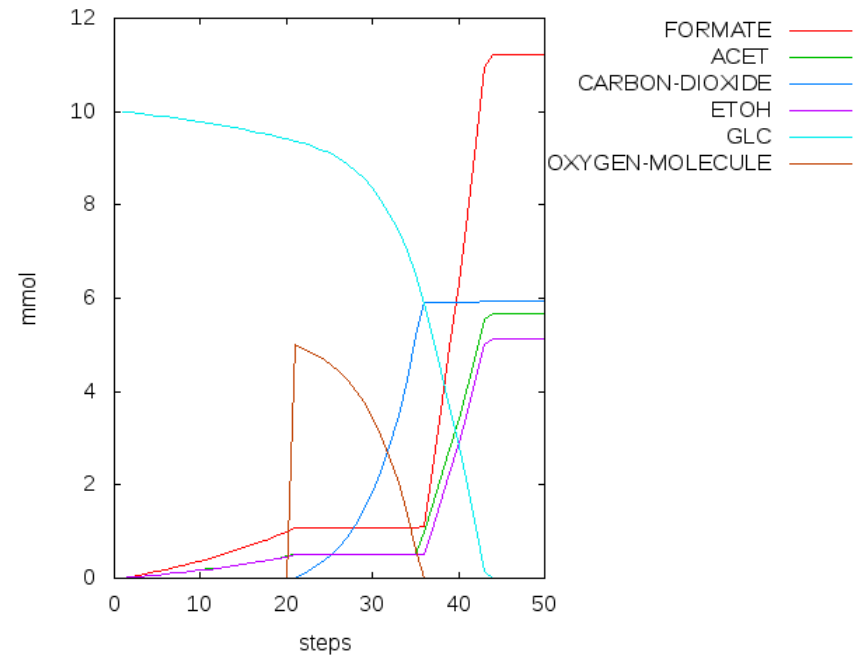
Painting *E. coli* Fluxes on Metabolic Map



Dynamic FBA Modeling of *E. coli*

- Dynamic FBA modeling of *E. coli* growth under varying nutrient conditions
  - t=1-20: E. coli grows anaerobically on 10 mmol glucose
  - t=21-34: O_2 is added to the simulation; E. coli grows completely aerobically
  - t=34-35: O_2 availability becomes limiting; acetate forms
  - t=36-44: O_2 is exhausted; anaerobic growth resumes
  - t=45 onwards: glucose is exhausted, cells begin to die





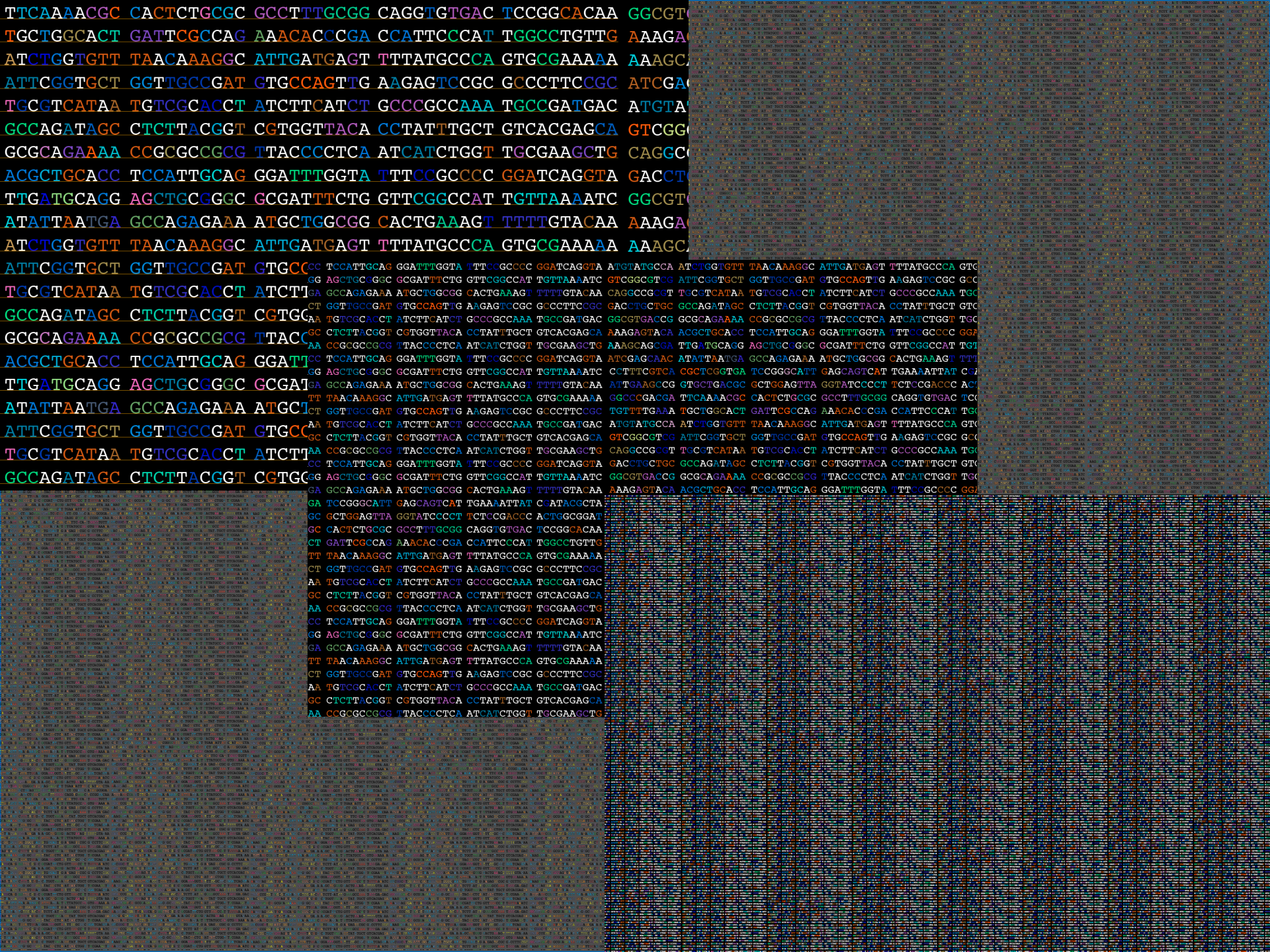
Where do Metabolic Models Come From?

A Bioinformatics Triumph: Automatic Construction of (Low Quality) Metabolic Models from Sequenced Genomes

Given: Genome sequence

Compute: Reaction network, nutrients, biomass metabolites

This



Human Genome Project

- Early publications about the Human Genome Project were remarkably vague

- How would the genes be found?
- What fraction of human genes would we find?
- How would we predict the functions of the genes?
- What fraction of genes would we predict functions for?
  - Today we can predict functions for ~50% of bacterial genes

- Tens of thousands of genomes sequenced to date
  - ~1,000 from human microbiome

The Solution

This

Gene Finding

This

What does a Bacterial Gene Look Like?

SD:6 8-10

ATCGGCTACCTGAATATGCACATATTATTCGAGCCGACA

AAGCTATGAATCCCTG…CCTTATAGACCTAGCTAATATA

GTG TAG

TTG TGA

Shine-Dalgarno sequence complementary to 16S rRNA

- Search upstream of ATG's for commonly found sequence element

- Create statistical model of within-gene sequence
- Accuracy: 90%

Bacterial Gene Finding

- Previous approach insufficient
  - Not all apparent “open reading frames” are real

- Use Markov chains to model the “flavor” of gene sequences
- A Markov chain is a sequence of variables X_i where the probability of X_i depends on the preceding k variables
- Models the probability of base b as depending on the k bases immediately before b in the sequence

- Accuracy: 90%

Predict Gene Function

This

One Approach

- Compare the sequence of protein P1 to the sequences of all other known proteins
  - Via the Genbank and UniProt databases – 23.6M proteins
- Find the protein P2 with the most similar sequences to P1
  - Using inexact string matching algorithm
  - In some cases, the function of P2 will have been determined experimentally
- Infer that the function of P1 is probably the same as that of P2

Human ATP Synthase
Aligned to E. coli ATP Synthase

```
>gnl|ECOLI|ATPA-MONOMER ATP synthase F1 complex - alpha subunit
          (complement(3917880..3916339)) Escherichia coli K-12
          substr. MG1655
         Length = 513

 Score =  541 bits (1393), Expect = e-155
 Identities = 290/513 (56%), Positives = 372/513 (72%), Gaps = 19/513 (3%)

Query: 48  TAEMSSILEERILGADTSVDLEETGRVLSIGDGIARVHGLRNVQAEEMVEFSSGLKGMSL 107
           + E+S ++++RI   +   +    G ++S+ DG+ R+HGL +    EM+        ++L
Sbjct: 5   STEISELIKQRIAQFNVVSEAHNEGTIVSVSDGVIRIHGLADCMQGEMISLPGNRYAIAL 64

Query: 108 NLEPDNVGVVVFGNDKLIKEGDIVKRTGAIVDVPVGEELLGRVVDALGNAIDGKGPIGSK 167
           NLE D+VG VV G    + EG  VK TG I++VPVG  LLGRVV+ LG  IDGKGP+
Sbjct: 65  NLERDSVGAVVMGPYADLAEGMKVKCTGRILEVPVGRGLLGRVVNTLGAPIDGKGPLDHD 124

Query: 168 TRRRVGLKAPGIIPRISVREPMQTGIKAVDSLVPIGRGQRELIIGDRQTGKTSIAIDTII 227
               V   APG+I R SV +P+QTG KAVDS++PIGRGQRELIIGDRQTGKT++AID II
Sbjct: 125 GFSAVEAIAPGVIERQSVDQPVQTGYKAVDSMIPIGRGQRELIIGDRQTGKTALAIDAII 184

Query: 228 NQKRFNDGSDEKKKLYCIYVAIGQKRSTVAQLVKRLTDADAMKYTIVVSATASDAAPLQY 287
           NQ+           + CIYVAIGQK ST++ +V++L +  A+  TIVV ATAS++A LQY
Sbjct: 185 NQR--------DSGIKCIYVAIGQKASTISNVVRKLEEHGALANTIVVVATASESAALQY 236

Query: 288 LAPYSGCSMGEYFRDNGKHALIIYDDLSKQAVAYRQMSLLLRRPPGREAYPGDVFYLHSR 347
           LAPY+GC+MGEYFRD G+ ALIIYDDLSKQAVAYRQ+SLLLRRPPGREA+PGDVFYLHSR
Sbjct: 237 LAPYAGCAMGEYFRDRGEDALIIYDDLSKQAVAYRQISLLLRRPPGREAFPGDVFYLHSR 296

Query: 348 LLERAAKMN----DAFGGG-------SLTALPVIETQAGDVSAYIPTNVISITDGQIFLE 396
           LLERAA++N    +AF  G       SLTALP+IETQAGDVSA++PTNVISITDGQIFLE
Sbjct: 297 LLERAARVNAEYVEAFTKGEVKGKTGSLTALPIIETQAGDVSAFVPTNVISITDGQIFLE 356

Query: 397 TELFYKGIRPAINVGLSVSRVGSAAQTRAMKQVAGTMKLELAQYREVAAFAQFGSDLDAA 456
           T LF  GIRPA+N G+SVSRVG AAQT+ MK+++G ++  LAQYRE+AAF+QF SDLD A
Sbjct: 357 TNLFNAGIRPAVNPGISVSRVGGAAQTKIMKKLSGGIRTALAQYRELAAFSQFASDLDDA 416

Query: 457 TQQLLSRGVRLTELLKQGQYSPMAIEEQVAVIYAGVRGYLDKLEPSKITKFENAFLSHVV 516
           T++ L  G ++TELLKQ QY+PM++ +Q  V++A  RGYL  +E SKI  FE A L++V
Sbjct: 417 TRKQLDHGQKVTELLKQKQYAPMSVAQQSLVLFAAERGYLADVELSKIGSFEAALLAYVD 476

Query: 517 SQHQALLGTIRADGKISEQSDAKLKEIVTNFLA 549
             H  L+  I   G  +++ + KLK I+ +F A
Sbjct: 477 RDHAPLMQEINQTGGYNDEIEGKLKGILDSFKA 509
```

Metabolic Model Generation

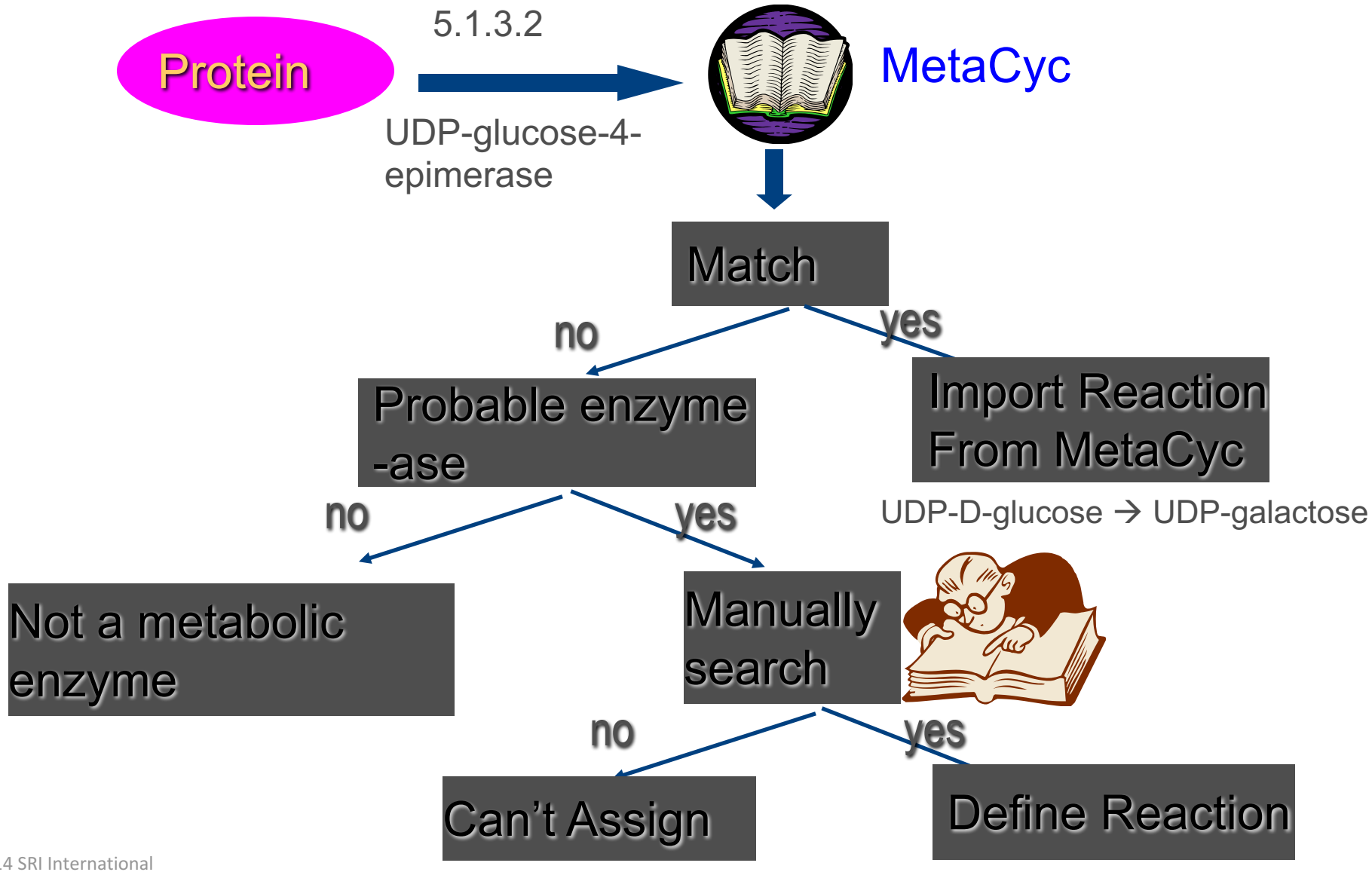
This

Reactome Inference

- For each protein in the organism, what reaction(s) does it catalyze?

- Protein (enzyme) activities can be specified in two ways
  - Enzyme names (uncontrolled vocabulary)
  - EC numbers (controlled vocabulary)

Match Enzymes to Reactions



MetaCyc: Curated Metabolic Database

	MetaCyc v20.1 2016	KEGG 2016	SEED 2015
Citations	51,000		
Pathways	2,500	320 Modules	583 Subsystems
Reactions	13,800	10,009	
Metabolites	13,400	17,554	
Mini-reviews (textbook pages)	7,500		

MetaCyc is free and open, contains computed atom mappings

"A Systematic Comparison of the MetaCyc and KEGG Pathway Databases *BMC Bioinformatics* 2013 14(1):112

Gap Filling of Metabolic Models

- Models created using preceding process are incomplete
- Use optimization methods to infer minimal number of reactions to import from MetaCyc to enable reachability of all biomass metabolites

- Experience from human metabolic model:
  - Gap filler proposed adding 8 new reactions from MetaCyc; 4 supported by literature research
  - Reversal of 4 reactions confirmed by literature searches
- From Bifidobacterium longum model:
  - +12 reactions proposed by gap filler
  - +13 reactions added during manual model development
  - 8 reactions in common

Summary of Further Research Needed

- Smarter more automated genome annotation
  - High levels of disagreement among genome annotation systems
- Smarter more automated metabolic model construction
- Develop suite of metabolic models for human microbiome organisms
- Improved visualization and analysis tools for metabolic models
- Improved interoperation of metabolic models

Acknowledgements

- SRI
  - **Suzanne Paley, Ron Caspi, Mario Latendresse, Ingrid Keseler, Carol Fulcher, Tim Holland, Markus Krummenacker, Tomer Altman, Richard Billington, Pallavi Kaipa**
- EcoCyc Collaborators
  - **Julio Collado-Vides, Robert Gunsalus, Ian Paulsen**
- MetaCyc Collaborators
  - **Lukas Mueller, Hartmut Foerster**
  - **Sue Rhee, Peifen Zhang**
- Funding sources:
  - **NIH National Institute of General Medical Sciences**

http://www.ai.sri.com/pkarp/talks/

BioCyc webinars:
biocyc.org/webinar.shtml