Microbiome Metabolic Modeling with Pathway Tools

Peter D. Karp
SRI International
Cost for High Quality Models of 300 Microbiome Organisms

• Multiple body sites of human, mouse, farm animals...
  – 500 species in human pan metagenome (Nielsen)
  – 321 organisms for human alone (Magnusdottir)

• 300 models x 2 months/model x $100K/yr = $5M
  – Assumes cost of combining models is zero

• If every model is duplicated 5 times: $25M
Challenges of Microbiome Metabolic Modeling

• **Scale**

• **Possible approaches**
  – Increase speed of manual model building
  – Collaborate / reuse organism models in community models
    • Understandability
    • Collaboration tools
  – Increase accuracy of automated gap filling
    • We'll still want to re-use high quality models

• **Reproducibility / Community Models as Reusable Tools**
  – How to disseminate tens/hundreds of coupled models?
Microbiome Modeling in Pathway Tools

• Key ideas:
  – Let's divide and conquer the problem of microbiome modeling
  – Plug-and-play metabolic models
  – Extensive and highly integrated software support for model development, interrogation, collaboration, reuse
    • We create bacterial models in 3-4 weeks
  – Marriage of systems biology and databases
    • Encode models as Pathway/Genome Databases (PGDBs)
Microbiome Modeling in Pathway Tools: Scale

- Speed manual development of individual models
  - Accurate qualitative metabolic reconstruction [minutes to days]
    - Based on MetaCyc metabolic database
    - Enzyme name matcher tool
    - Pathway prediction fills many reaction gaps
  - Reaction gap filler
  - Development mode identifies non-produced biomass metabolites
  - Identify base blocking metabolites and the reactions they block
  - Reaction balance checker
  - Fast interpretation of model results via visualization tools
  - Run models programmatically via PythonCyc API
    - Programming not required to build models with Pathway Tools
Microbiome Modeling in Pathway Tools: Collaborative Development / Reuse of Models

– Enhance model understandability
  • Data provenance via evidence codes, comments, citations, author credits
  • Couple models with enriching information
    – Chemical structures, atom mappings, pathways, genome, regulatory network
  • Web and desktop query and visualization tools: publish models on the web
  • Coming in 2016: run models through the web

– Collaborative model development
  • Pathway Tools enables concurrent multi-user updating of PGDBs
  • Transaction history stored for PGDBs
  • Editing tools: reaction editor, pathway editor, Marvin
  • Share PGDBs via PGDB registry -- http://biocyc.org/registry.html

– Plug-and-play models
  • All PGDBs share reaction and metabolite identifiers with MetaCyc
  • All PGDBs share common identifiers for cellular compartments
Pathway Tools Enables Multi-Use Metabolic Databases

- Encyclopedia
- Metabolic Model
- Queryable Database
- Zoomable Metabolic Map
- Omics Data Analysis
Pathway Tools Software: Workflow

- Annotated Genome + PathoLogic
- MetaFlux
- Pathway/Genome Database
- Pathway/Genome Editors
- Pathway/Genome Navigator

640,000 lines of Lisp code ≈ 1.5M lines of C or Java code

© 2014 SRI International
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 metabolic network (qualitative metabolic reconstruction)
  – Reactome, metabolites, metabolic pathways imported from MetaCyc
  – Batch mode for high throughput
• Predicts which genes code for missing enzymes in metabolic pathways
• Infers transport reactions from transporter names in genome
Metabolic Model Generation in Pathway Tools

• PathoLogic: Qualitative metabolic model reconstruction
  – Inference of the reactome from the annotated genome
  – Inference of metabolic pathways by selecting from MetaCyc pathways

• MetaFlux
  – Infer biomass reaction
  – Gap fill
  – Edit reaction complement
  – Iterate
  – Karp et al, *Briefings in Bioinformatics* 2015, in press
Pathway Prediction

• Pathway prediction is useful because
  – Pathways organize the metabolic network into mentally tractable units
  – Pathways can be used for analysis of computed fluxes and high-throughput data
    • Visualization, enrichment analysis
  – Pathway inference fills gaps in metabolic network
    • Reduces computational demands of gap filling
Reactome Inference

• For each protein in the organism, infer reaction(s) it catalyzes

• Build from existing genome annotation!

• Match protein functions to MetaCyc reactions
  – Enzyme names  (uncontrolled vocabulary)
  – EC numbers
  – Gene Ontology terms
PathoLogic Enzyme Name Matcher

- Name matcher generates alternative variants of each name and matches each to MetaCyc
- Strips extraneous information found in enzyme names

- Putative carbamate kinase, alpha subunit
- Flavin subunit of carbamate kinase
- Cytoplasmic carbamate kinase
- Carbamate kinase (abcD)
- Carbamate kinase (3.2.1.4)
MetaCyc: Curated Metabolic Database

<table>
<thead>
<tr>
<th></th>
<th>MetaCyc v19.0 2015</th>
<th>KEGG 2013</th>
<th>SEED 2015</th>
</tr>
</thead>
<tbody>
<tr>
<td>Citations</td>
<td>45,000</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Pathways</td>
<td>2,310</td>
<td>179 Modules</td>
<td>583 Subsystems</td>
</tr>
<tr>
<td>Reactions</td>
<td>12,400</td>
<td>8,692</td>
<td></td>
</tr>
<tr>
<td>Metabolites</td>
<td>12,000</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Mini-reviews:</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Textbook Pages</td>
<td>6,300</td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

- MetaCyc is free and open
- Contains large number of computed atom mappings

“A Systematic Comparison of the MetaCyc and KEGG Pathway Databases
BMC Bioinformatics 2013 14(1):112
Algorithm for Inference of Metabolic Pathways

• For each pathway in MetaCyc consider
  – What fraction of its reactions are present in the organism's reactome?
  – Are enzymes present for reactions unique to the pathway?
  – Is a given pathway outside its designated taxonomic range?
    • Calvin cycle: green plants, green algae, etc
  – Are enzymes present for designated “key reactions” within MetaCyc pathways?
    • Calvin cycle / ribulose bisphosphate carboxylase

Standards in Genomic Sciences 5:424-429 2011
## Pathway Evidence Report

### Biosynthesis: Cofactors, Prosthetic Groups, Electron Carriers

<table>
<thead>
<tr>
<th>Pathway</th>
<th>Pathway Glyph</th>
<th>Total Rxns</th>
<th>Rxns Present in V. cholerae</th>
<th>Rxns Present in Other Pwys</th>
<th>Other Pwys</th>
</tr>
</thead>
<tbody>
<tr>
<td>biosynthesis of proto- and archeine</td>
<td></td>
<td>15</td>
<td>12</td>
<td>2</td>
<td>(FRDA charging pathway) cobalamin biosynthesis II (aerobic pathway)</td>
</tr>
<tr>
<td>biotin biosynthesis I</td>
<td></td>
<td>4</td>
<td>4</td>
<td>0</td>
<td>(none)</td>
</tr>
<tr>
<td>cobalamin biosynthesis I</td>
<td></td>
<td>7</td>
<td>6</td>
<td>2</td>
<td>cobalamin biosynthesis II (aerobic pathway)</td>
</tr>
<tr>
<td>cobalamin biosynthesis II (aerobic pathway)</td>
<td></td>
<td>18</td>
<td>5</td>
<td>3</td>
<td>cobalamin biosynthesis I biosynthesis of proto- and archeine</td>
</tr>
</tbody>
</table>

© 2014 SRI International
Pathway Tools Software: Pathway/Genome Editors

- Interactively update PGDBs with graphical editors
- Support geographically distributed teams of curators with object database system
- Gene and protein editor
- Reaction editor
- Compound editor
- Pathway editor
- Operon editor
- Publication editor
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***
  – Cellular-compartment aware
  – Removal of flux loops, inference of biomass reaction

• **Knock-out mode** (single/double gene/reaction knock-outs)

• **Model development mode**
  – Development mode (multiple gap filling)
  – Fast Development mode (reaction gap filling)  [Latendresse 2014]
  – Identify dead-end metabolites and **blocked reactions***

Solver Used by MetaFlux

- MetaFlux formulates LP and MILP problems for SCIP solver and processes SCIP output
  - MILP is fast
  - Multi-core version under development
  - Free to academics, distributed as part of Pathway Tools
  - Konrad-Zuse-Zentrum für Informationstechnik Berlin
  - scip.zib.de
MetaFlux Multiple Gap Filling of FBA Models

• Reaction gap filling (Kumar et al, BMC Bioinf 2007 8:212):
  – Reverse directionality of selected reactions AND
  – Add minimal set of reactions from MetaCyc to model to enable a solution
  – Reaction cost is a function of reaction taxonomic range
  – Compartment-based gap filling available
  – FastGap gap filler performs reaction gap filling only, but is much faster

• Partial solutions: Identify maximal subset of biomass components for which model can yield positive fluxes

• Metabolite gap filling: Postulate additional nutrients and secretions
Other Aspects of MetaFlux

- Store and update metabolic model within PGDB
  - Curate metabolic model and organism database simultaneously
  - All query and visualization tools applicable to FBA model

- Graphical exploration of model results
  - Visualize computed reaction fluxes using cellular overview
  - Plot organism composition, metabolite concentrations

- Programmatic execution of models from Lisp and PythonCyc APIs
  - Other APIs exist but don't currently support modeling
    - Perl, Java, R
**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
*E. coli* Fluxes on Pathway Diagram

**Escherichia coli** K-12 subcell. MG1655 Pathway: S-adenosyl-L-methionine cycle | [More Detail | Less Detail | Species Comparison]

- 7.77e-4
  - methionine
  - adenosyltransferase: metK
  - 2.5.1.6
- 0.0874
  - cobalamin-independent homocysteine
  - transmethyllase: metE
  - methionine synthase: metH
  - 2.1.1.14
- tetrahydropteroyl tri-L-glutamate
- diphosphate phosphate
- L-methionine
- S-adenosyl-L-methionine
- a demethylated methyl donor
- 2.1.1-
- a methylated methyl donor
- S-adenosyl-L-homocysteine
- S-adenosylhomocysteine nucleosidase: mtn
  - 3.06e-4
- S-ribosyl-L-homocysteine
- tetrahydropteroyl tri-L-glutamate
- S-ribosylhomocysteine lyase: luxS
  - 4.41.21
- autoinducer 2
- autoinducer AI-2 biosynthesis I

© 2014 SRI International
Dynamic FBA Modeling of \textit{E. coli}

- Dynamic FBA modeling of \textit{E. coli} growth under varying nutrient conditions
  - \( t=1-20 \): \textit{E. coli} grows anaerobically on 10 mmol glucose
  - \( t=21-34 \): \( \text{O}_2 \) is added to the simulation; \textit{E. coli} grows completely aerobically
  - \( t=34-35 \): \( \text{O}_2 \) availability becomes limiting; acetate forms
  - \( t=36-44 \): \( \text{O}_2 \) is exhausted; anaerobic growth resumes
  - \( t=45 \) onwards: glucose is exhausted, cells begin to die
Dynamic Grid Modeling of a Simple Microbial Community

Inspired by Segre's COMETS

- Initially, *E. rectale* is present throughout the grid; *E. coli* is present in southwest corner
- Halfway through simulation, *B. thetaiotaomicron* is added to the middle of the lawn
- *E. rectale* shows higher growth where *E. coli* or *B. theta* are present because of availability of acetate from *E. coli*. *E. rectale* produces butyrate where acetate is present.
Curated Models in Hand

- *Escherichia coli*
- *Eubacterium rectale*
- *Bacteriodes thetaiotamicron*
- *Methanobrevibacter smithii* (in progress)

- Will be present in BioCyc.org and PGDB registry in November

- Contact me to coordinate on creation of additional models
- Biocyc.org/microbiome-modeling.shtml

- Attend metabolic modeling tutorials at SRI (see BioCyc.org)
PythonCyc: The Python API for Pathway Tools

Pathway Tools API calls in Python

```python
# Import PythonCyc plus a few other libraries
import matplotlib
import pythoncyc as pc
import numpy as np
import matplotlib.pyplot as plt

# Select the EcoCyc PDB
eco = pc.select_organism('ecoli')

# Issue queries to Pathway Tools in Python!
print "PythonCyc Demo #1: Hello World"
print "number of small molecule metabolic reactions: {0}.format(len(eco.all_reactions('ssm')))
print "number of reactions, total: {0}.format(len(eco.all_reactions('all')))
```

PythonCyc Demo #1: Hello World
number of small molecule metabolic reactions: 1555
number of reactions, total: 2463

Pathway Tools and MetaFlux plots with *matplotlib*

Get PythonCyc at:
https://github.com/latendre/PythonCyc
MetaFlux JSON output: *cobrapy* meets MetaFlux

Load, manipulate MetaFlux models in *cobrapy*

```python
import cobra

cobra_metaflux = cobra.io.load_json_model("/Users/weaver/projects/pynb/ecoli-19.5-om-aer.json")
cobra_metaflux.objective = "Max growth"
cobra_metaflux.optimize()

cobra_metaflux_aner = cobra.io.load_json_model("/Users/weaver/projects/pynb/ecoli-19.5-om-aer.json")
cobra_metaflux_aner.objective = "Max growth"
cobra_metaflux_aner.optimize()

cobra_metaflux = cobra.io.load_json_model("/Users/weaver/projects/pynb/ecoli-19.5-om-aer.json")
cobra_metaflux.objective = "Max growth"
cobra_metaflux.optimize()

cobra_metaflux.aner = cobra.io.load_json_model("/Users/weaver/projects/pynb/ecoli-19.5-om-aer.json")
cobra_metaflux.aner.objective = "Max growth"
cobra_metaflux.aner.optimize()
```

Snippet from EcoCyc-19.5-GEM JSON

```json
{
  "name": "Ecoli_Metaflux_GEM",
  "description": "Pathway Tools MetaFlux-create",
  "version": "19.5",
  "compartments": {
    "p": "periplasm",
    "o": "exacellular",
    "l": "inner membrane",
    "o": "outer membrane",
    "e": "cytoplasm"
  },
  "reactions": {
    "gene_reaction_rule": "EG10592 and EG10594",
    "id": "TRANS_RXNO 541 L2R",
    "upper_bound": 0.000000,
    "lower_bound": 30000.000000,
    "name": "TRANS-RXNO-541",
    "metabolites": {
      "Pi_c": 1.000000,
      "PROTON_c": 1.000000,
      "ATP_c": -1.000000,
      "WATER_c": -1.000000,
      "ADP_c": 1.000000,
      "METHYL_BETA_D_GALACTOSIDE_c": 1.000000,
      "METHYL_BETA_D_GALACTOSIDE_p": -1.000000
    }
  },
  "name": "Ecoli_Metaflux_aner",
  "description": "Pathway Tools MetaFlux-create",
  "version": "19.5",
  "compartments": {
    "p": "periplasm",
    "o": "exacellular",
    "l": "inner membrane",
    "o": "outer membrane",
    "e": "cytoplasm"
  },
  "reactions": {
    "gene_reaction_rule": "EG10592 and EG10594",
    "id": "TRANS_RXNO 541 L2R",
    "upper_bound": 0.000000,
    "lower_bound": 30000.000000,
    "name": "TRANS-RXNO-541",
    "metabolites": {
      "Pi_c": 1.000000,
      "PROTON_c": 1.000000,
      "ATP_c": -1.000000,
      "WATER_c": -1.000000,
      "ADP_c": 1.000000,
      "METHYL_BETA_D_GALACTOSIDE_c": 1.000000,
      "METHYL_BETA_D_GALACTOSIDE_p": -1.000000
    }
  }
}
```

Works with any MetaFlux model (*E. rectale* here)

```python
# Create cobrapy-compatible JSON models with any PGDB
# Here I use Subacterium rectale, a common gut Firmicutes
subacterium = cobra.io.load_json_model("/Users/weaver/projects/pynb/erect-19.5-autogen-draft.json")
subacterium.objective = "Max growth"
subacterium.optimize()

Conditions: Subacterium rectale anaerobic growth on glucose (2 mmol/gDw/hr) w/ acetate
```

MetaFlux/SCIP E. rectale growth rate: 0.093815
MetaFlux/cobrapy E. rectale growth rate: 0.093815
Acknowledgements

• Mario Latendresse
  – MetaFlux implementation
  – PythonCyc API

• Dan Weaver
  – Model building
  – PythonCyc API
  – Cobrapy interface

• MetaCyc Collaborators
  – Sue Rhee, Peifen Zhang
  – Lukas Mueller, Hartmut Foerster

Funding sources:
– NIH National Institute of General Medical Sciences

BioCyc webinars: biocyc.org/webinar.shtml

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