Integrative Pathway Mapping

Sara Calhoun
Andrej Sali Lab, UCSF

Enzyme Function Initiative (EFI)
2015 ASBMB Meeting
March 30, 2015
Integrative pathway mapping

1. Prediction of the function in the context of the pathway
2. Prediction of the pathway based on all available information
3. Automated (i.e., thorough, objective, ...) method

What is the pathway, given a set of potential enzymes and metabolites as well as at least one enzyme and/or metabolite in it?
A tiny subset of structure-, imaging- and systems-based information that (in principle) informs network prediction

**Structural biology**
- Virtual screening (VS)
- *In silico* transformations and chemical similarity calculations
- High-throughput screening

**Systems biology and imaging**
- Coexpression of genes
- Genome context
- Genetic interactions
- Metabolite levels
- Protein functional links
- Orthology
- Tomography
- Super-resolution optical microscopy
- FRET spectroscopy

---

**Diagram**

**Protein interaction predictions**
- Virtual screening
- Cheminformatics
- Gene coexpression
- High-throughput screening

**Genome neighborhood**
- Proteomics
- Chemogenomics
- Metabolomics
- Regulons
Mindset: From integrative structural biology towards modeling networks and pathways

Components
- Sampling and scoring
- Cross-linking
- Restraints
- Co-purification
- Docking
- Bioinformatics, physics
- Electron microscopy
- NMR
- X-ray crystallography
- Small angle X-ray scattering
- Proteomics
- HDXMS
- Proteomics
- Co-purification
- Bioinformatics, physics

Alber et al. Annual Reviews in Biochemistry, 2008
Mindset: From integrative structural biology towards modeling networks and pathways
Pathway model represented as a graph

- Enzymes and ligands are individual nodes
- Edges reflect a relationship between an enzyme and its substrate/product
Scoring

- Translation of information into restraints that quantify the degree of consistency between a pathway model and the corresponding information
- Scoring function is a sum of individual restraints to rank alternative models

\[ Z_{comp} = Z_{sea} + Z_{rxns} + Z_{dock} + Z_{screen} \]
Virtual screening

- Computational docking methods predicts the relative affinity of substrate-enzyme interactions relative to all other metabolites included in screening library.
- When the structure is unavailable, homologs or models can be used.

\[
Z_{\text{dock}} = \frac{1}{N} \sum_{i=1}^{N} Z_i
\]

<table>
<thead>
<tr>
<th>Metabolite</th>
<th>Score</th>
<th>Z-Score</th>
</tr>
</thead>
<tbody>
<tr>
<td><img src="image1" alt="Chemical Structure" /></td>
<td>-10</td>
<td>2.5</td>
</tr>
<tr>
<td><img src="image2" alt="Chemical Structure" /></td>
<td>-9</td>
<td>2.0</td>
</tr>
<tr>
<td><img src="image3" alt="Chemical Structure" /></td>
<td>-6</td>
<td>1.8</td>
</tr>
<tr>
<td><img src="image4" alt="Chemical Structure" /></td>
<td>-2</td>
<td>0.6</td>
</tr>
<tr>
<td>...</td>
<td>...</td>
<td>...</td>
</tr>
</tbody>
</table>
In silico chemical transformations

Chemical similarity between transformed substrate and product

- Transformations encoded by SMARTS patterns and applied \textit{in silico}
- Compounds represented as Morgan fingerprints and compared by the Tanimoto coefficient

\[
Tanimoto \text{ coefficient} \rightarrow Z_{1,2,E} \\
Z_{rxn} = \frac{1}{N} \sum_{i} Z_{i}
\]

Software: OpenEye, RDKit
Similarity ensemble analysis

- SEA, a similarity metric between ensembles of ligands, can be used to predict functional relationships between proteins.

\[
S_{\text{pathway}} = \frac{1}{N} \sum_{i}^{N} S_i \rightarrow Z_{sea}
\]

M. Keiser, B. Shoichet
Sampling: Monte Carlo with simulated annealing

1) Starting with a model, select a node to alter identity

2) Select a new node identity from all possible node identities & make change in model

3) If change improves score, accept the new model. Otherwise, accept the new model with a probability determined by the difference in scores.

4) Repeat until "convergence"

\[ \Delta S = \text{Score(model}_{i+1}) - \text{Score(model}_i) \]

\[ \text{Prob. of acceptance} = \exp(-\Delta S/T) \]
<table>
<thead>
<tr>
<th>Biochemical Pathway</th>
<th>Number of Enzymes</th>
<th>Number of Metabolites</th>
</tr>
</thead>
<tbody>
<tr>
<td>Glycolysis, <em>E. coli</em></td>
<td>10</td>
<td>2,965</td>
</tr>
<tr>
<td>KDO-8P pathway</td>
<td>4</td>
<td>3,336</td>
</tr>
<tr>
<td>Serine/cysteine biosynthesis</td>
<td>5</td>
<td>3,494</td>
</tr>
</tbody>
</table>
Sampling of pathways, starting with random initial guesses, saturates the space of good scoring pathways.

Clustering of pathway models using distance calculated from number of mismatched assignments.
Assessment: synergy of different types of information

Glycolysis

<table>
<thead>
<tr>
<th>Step in pathway</th>
<th>Rank by VS score alone</th>
<th>Rank using VS scores and SEA</th>
<th>Rank using chemical transformations and SEA</th>
<th>Rank using VS, chemical transformations, and SEA</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>37</td>
<td>229</td>
<td>6</td>
<td>1</td>
</tr>
<tr>
<td>2</td>
<td>108</td>
<td>189</td>
<td>9</td>
<td>1</td>
</tr>
<tr>
<td>3</td>
<td>18</td>
<td>25</td>
<td>7</td>
<td>1</td>
</tr>
<tr>
<td>4</td>
<td>183</td>
<td>323</td>
<td>3</td>
<td>1</td>
</tr>
<tr>
<td>5</td>
<td>12</td>
<td>125</td>
<td>1</td>
<td>1</td>
</tr>
<tr>
<td>6</td>
<td>117</td>
<td>24</td>
<td>1</td>
<td>1</td>
</tr>
<tr>
<td>7</td>
<td>71</td>
<td>195</td>
<td>1</td>
<td>1</td>
</tr>
<tr>
<td>8</td>
<td>5</td>
<td>107</td>
<td>1</td>
<td>1</td>
</tr>
<tr>
<td>9</td>
<td>27</td>
<td>10</td>
<td>1</td>
<td>1</td>
</tr>
<tr>
<td>10</td>
<td>129</td>
<td>114</td>
<td>1</td>
<td>1</td>
</tr>
</tbody>
</table>
Benchmark

<table>
<thead>
<tr>
<th>Biochemical Pathway</th>
<th>Number of Possible Pathways</th>
<th>Rank of Correct Pathway</th>
</tr>
</thead>
<tbody>
<tr>
<td>Glycolysis, <em>E. coli</em></td>
<td>$5 \times 10^{44}$</td>
<td>1</td>
</tr>
<tr>
<td>KDO-8P pathway</td>
<td>$9 \times 10^{18}$</td>
<td>18</td>
</tr>
<tr>
<td>Serine/cysteine biosynthesis</td>
<td>$2 \times 10^{23}$</td>
<td>1</td>
</tr>
</tbody>
</table>

Prediction of the enzyme order and their metabolites is perfect or nearly so.

Benchmarking of the prediction of the enzymes in the pathway, their order and metabolites is “satisfactory”, even when “dummy” enzymes and metabolites need to be used.
Prediction of pathway downstream of \textit{H. influenzae} SBP TRAP transporter

Tripartite ATP-independent periplasmic transporters - family of transporters in bacteria and archaea for the uptake of organic acids

\textbf{Haemophilus influenzae}

“Bacterial flu” found in the upper respiratory system of host

Centers for Disease Control and Prevention

Andrej Sali  
Brian Shoichet  
Magdalena Korczynska  
Matthew Jacobson  
Suwen Zhao  
John Gerlt  
Daniel Wichelecki  
Brian San Francisco

Steven Almo  
Matthew Vetting  
Nawar Al-Obaidi  
Andrei Osterman  
Dmitry Rodionov
Solute binding protein TRAP in *H. influenzae*

Thermofluor screening hits against *Haemophilus influenzae* RdAW SBP TRAP

L-gulonate  
(\( \Delta T_m = 10.9^\circ C \))

D-mannonate  
(\( \Delta T_m = 7.3^\circ C \))

Same Thermofluor screening hits against *M. haemolytica* TRAP

Screening by NYSGXRC
Thermofluor assay hits

Scoring term is derived from maximum Tanimoto coefficient between a given ligand and hits from thermofluor assay

\[ Tanimoto \text{ coefficient} \rightarrow Z_{\text{screen}} \]
Genome context & comparative genomics of TRAP transporter in *H. influenzae*

From Suwen Zhao, Jacobson Lab

**Regulon of UxuR**

RegPrecise Database

based on transcription factor binding site motifs

<table>
<thead>
<tr>
<th>gene</th>
<th>annotation</th>
</tr>
</thead>
<tbody>
<tr>
<td>uxuR</td>
<td>hexuronate utilization operon repressor</td>
</tr>
<tr>
<td>uxuA</td>
<td>mannonate dehydratase</td>
</tr>
<tr>
<td>HI0053</td>
<td>oxidoreductase</td>
</tr>
<tr>
<td>uxuQ</td>
<td>predicted glucuronate TRAP</td>
</tr>
<tr>
<td>uxuP</td>
<td>predicted hexuronate TRAP</td>
</tr>
<tr>
<td>uxuM</td>
<td>predicted hexuronate TRAP</td>
</tr>
<tr>
<td>kdgK</td>
<td>2-dehydro-3-deoxygluconate kinase</td>
</tr>
<tr>
<td>uxuB</td>
<td>mannonate oxidoreductase</td>
</tr>
<tr>
<td>eda</td>
<td>2-dehydro-3-deoxyphosphogluconate aldolase</td>
</tr>
</tbody>
</table>

**Pfams conserved in neighborhood of TRAP homologs across different bacterial genomes**
Docking and modeling summary

16,347 KEGG ligands using DOCK against each of six protein models

**C9MHP1**
dehydrogenase

33% sequence identity (SID) to template

**C9MHP2**
transporter SBP

27% SID

**C9MHP5**
kinase

44% SID

**C9MHP6**
reductase

34% SID

**C9MHP7**
aldolase

Crystal structure

**C9MHN9**
dehydratase

74% SID

Docking by Magdalena Korczynska and Brian Shoichet
Scoring term is derived from maximum Tanimoto coefficient between a given ligand and central metabolism compounds
Virtual screening
Chemical transformations
Gathering information
Genome neighborhood
High-throughput screening
Monte Carlo simulated annealing
Assessing accuracy
Assessing sampling
Gathering models
Sampling good models
Designing model representation and evaluation
Pathway model
Potential ligands
Potential proteins
6 protein-ligand pairs
5 enzymatic reactions
1 protein screened
1 endpoint
6 proteins
6 proteins
6 proteins
6 proteins
Similarity ensemble approach
Assessing accuracy
Assessing sampling
TRAP pathway sampling

Out of $2 \times 10^{23}$ possible pathways, 154 good-scoring pathway models grouped into 12 clusters.

Example of one MC run

Matches best scoring pathway

```
Step 0
0 ----> 3 ----> 4 ----> 1 ----> 2 ----> 5
```
Pathway predictions

Representatives from top five clusters

<p>| | | | | | | |</p>
<table>
<thead>
<tr>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td>0</td>
<td>1</td>
<td>2</td>
<td>3</td>
<td>4</td>
<td>5</td>
<td></td>
</tr>
<tr>
<td>transporter SBP C9MHP2</td>
<td>dehydrogenase C9MHP1</td>
<td>reductase C9MHP6</td>
<td>dehydratase C9MHN9</td>
<td>kinase C9MHP5</td>
<td>aldolase C9MHP7</td>
<td></td>
</tr>
</tbody>
</table>
SBP TRAP pathway prediction

**HiGulDH**

L-gulonate → D-fructuronate → D-mannonate

**HiGulPQM**

periplasm

**HiFR**

D-fructuronate → D-mannonate

UxuA

**HiKdgpA**

D-glyceraldehyde-3P

2-keto-3-deoxy-6P-D-gluconate

**HiKdgK**

2-keto-3-deoxy-D-gluconate

**HiKdgpA**

central carbon metabolism

in vitro assays by Daniel Wichlecki, Gerlt lab

Transcripts and growth assays by Brian San Francisco, Gerlt lab

Structure by Matthew Vetting et al., NYSGXRC
Validated SBP TRAP pathway prediction

periplasm

L-gulonate \( \rightarrow \) D-fructuronate \( \rightarrow \) D-mannonate

\( \text{HiGulDH} \)

\( k_{\text{cat}}/K_m = 9.3 \times 10^3 \text{ M}^{-1}\text{s}^{-1} \)

\( \text{HiFR} \)

\( 7.6 \times 10^5 \text{ M}^{-1}\text{s}^{-1} \)

\( \text{HiUxuA} \)

\( 1.1 \times 10^3 \text{ M}^{-1}\text{s}^{-1} \)

\( 27\text{-fold} \)

\( \text{HiKdgpA} \)

\( 7.5 \times 10^3 \text{ M}^{-1}\text{s}^{-1} \)

\( 13\text{-fold} \)

\( \text{HiKdgK} \)

\( 2.4 \times 10^4 \text{ M}^{-1}\text{s}^{-1} \)

\( 28\text{-fold increase in expression on gulonate vs. glucose} \)

\( \text{2-keto-3-deoxy-6P-D-gluconate} \)

\( \text{2-keto-3-deoxy-D-gluconate} \)

D-glyceraldehyde-3P

\( \text{PDB id: 4PBQ} \)

10- to 17-fold

no growth KO

central carbon metabolism

in vitro assays by Daniel Wichlecki, Gerlt lab

Transcripts and growth assays by Brian San Francisco, Gerlt lab

Structure by Matthew Vettering et al., NYSGXRC
Annotation of enzyme function

1. Prediction of the function in the context of the pathway
2. Prediction of the pathway based on all available information
3. Automated (*ie*, thorough, objective, …) method
4. We’ve demonstrated it’s performance on retrospective test cases
5. We’ve generated predictions and validated a previously uncharacterized pathway
6. Future development and applications

*Integrative Modeling Platform (IMP)*


Open source, versions, documentation, wiki, examples, mailing lists, unit testing, bug tracking, …
Acknowledgements

Andrej Sali
  Daniel Russel
  Martin Steinegger
  Ben Webb

Brian Shoichet
  Magdalena Korczynska
  Matthew O’Meara
  Henry Lin

Matt Jacobson
  Suwen Zhao
  Chakrapani Kalyanaraman

Patsy Babbitt
  Shoshana Brown

Andrei Osterman
  Dmitry Rodionov

John Gerlt
  Daniel Wichelecki
  Brian San Francisco

Steven Almo / NYSGXRC
  Matthew Vetting
  Nawar Al-Obaidi

Enzyme Function Initiative labs

Karen Allen