Computational Inference of Metabolic and Regulatory Networks in Bacteria

7/19/2007
The collection of all the metabolic pathways in an organism comprise its metabolic networks that generate essential components such as amino acids, sugars and lipids, and the energy required to synthesize them and to use them in creating proteins and cellular structures.
Regulatory networks are another kind of network that are of central importance in biology, which sense the intracellular and extracellular environmental changes, transmit the resulting signal into the cell, and orchestrate the cellular functions according to the environmental changes.

http://www.avatar.se/strbio2001/metabolic/eqf.gif/
Regulatory Networks

- If the regulation of genes are involved, the regulatory networks are often called genetic networks.

- Typically, metabolic and regulatory networks are viewed as different entities. In metabolic networks the flow of mass and energy is the essential purpose of the machinery.

- In regulatory networks the purpose is the regulation of other processes, and the use of energy and mass flow is a requirement, but not really the point.

- However, there is an essential component of regulation also in metabolic networks: The enzymes are regulated through interactions with substrates and products so that the appropriate conditions in the cell are upheld.
Computational prediction of metabolic and regulatory networks in bacteria

- Traditionally, metabolic and regulatory networks are characterized by ad hoc experimental efforts; and we have a good understanding of them in some model organisms.

- However, it is not possible to study the metabolic and regulatory networks in all organisms by experiments.

- The availability of genome sequences have provided an unprecedented opportunities for computationally inferring these networks in less studied sequenced genomes using comparative genomics as well as other approaches.

- Computational studies should also be able to provide some insights into the unique parts of the networks in a genome.
Computational prediction of metabolic and regulatory networks in bacteria

- Metabolic and regulatory network prediction problem: give genomic sequence of an organism and some models organisms, known metabolic and regulatory networks in these models organisms, and other public databases, predict the metabolic and regulatory networks in the target genome.

- Choice of model organisms: should be as phylogenetically close as possible;

- Public databases can be used: microarray databases; protein-protein interaction databases; knock-out/down phenotype databases; etc.
An example of network prediction: prediction of networks in *Synechococcus sp. WH 8102*

- A member of cyanobacteria
- Ubiquitously distributed in World’s oceanic regions
- A major contributor of global primary production of $\text{CO}_2$ fixation

**WH8102 genome**

Palenik, B. et. al. *Nature*, 2003, **424**, 1037-1042

Size: 2.4 Mbp
ORFs: 2516
Response regulators: 46

Size: 2.4 Mbp
ORFs: 2516
Response regulators: 46
Objectives

1. To predict response regulatory networks of *Synechococcus sp.* WH8102, including
   - **Nutrients assimilation networks**
     - Phosphorus assimilation networks
     - Nitrogen assimilation networks
     - Iron assimilation networks
     - Carbon fixation networks
     - .......
   - **Photosynthesis acclimation regulatory networks**
     - Light quality acclimation
     - Light quantity acclimation
   - **Osmo-regulatory networks**
   - **Circadian regulatory networks**
Objectives

2. To predict gene transcription regulatory networks in WH 8102

A partial gene regulatory network in *E. coli*

The overall strategies

- **Template-based pathway/networks inference protocol**
  Rationale: Genes and regulatory pathways are more or less conserved in evolutionarily related organisms

  Step 1: Derive “homologous” template pathways/networks in related genomes through mining databases and literatures

  Step 2: Map template pathways to the target genome to derive sub-pathways

  Step 3: Integrate the sub-pathways to derive an initial pathway/network

  Step 4: Expand the initial pathway to construct a temporary working model by recruiting additional genes

  Step 5: Experimental validation and refinement --- working model

Construction of template pathways/networks in related genomes through mining databases and literatures

Nitrogen assimilation pathways are relatively well studied in

- *Nostoc sp.* PCC 7120
- *Synechocystis* PCC 6803
- *Synechococcus sp.* PCC 7942
- *Synechococcus sp.* WH 8103
Construction of template nitrogen assimilation networks
--- an example in Nostoc sp. PCC 7120

Plasma membrane

\[ \text{GlnB} \xrightarrow{\text{kinase}} \text{GlnB-p} \]

\[ \text{GlnB-p} \xrightarrow{\text{phosphatase}} \text{GlnB} \]

\[ \text{CO}_2 \text{fixation} \]

\[ \text{isocitrate} \xrightarrow{lcd} \text{2-oxglutarate} \]

\[ \text{GlnA} \xrightarrow{\text{glutamine}} \text{NtcA} \]

\[ \text{NtcA} \xrightarrow{\text{NifK}} \text{NifE} \]

\[ \text{NifE} \xrightarrow{\text{NifN}} \text{NifX} \]

\[ \text{NifX} \xrightarrow{\text{orf2}} \text{nifBW} \]

\[ \text{nifBW} \xrightarrow{\text{hesA}} \text{hesB} \]

\[ \text{fdxH} \xrightarrow{\text{orf3}} \text{mop} \]

\[ \text{hetC} \xrightarrow{\text{petH}} \text{devB} \]

\[ \text{devB} \xrightarrow{\text{devC}} \text{devA} \]

\[ \text{NO}_3^-/\text{NO}_3^{2-} \]

\[ \text{NirA} \xrightarrow{\text{NarB}} \text{NH}_4^+ \]

\[ \text{Urea} \xrightarrow{\text{Urea}} \text{UrtABCD} \]

\[ \text{N}_2 \text{fixation} \]

\[ \text{NH}_4^+ \]

\[ \text{UrtABCD} \]

\[ \text{Amt} \]

\[ \text{NO}_3^- \]

\[ \text{NO}_2^- \]

\[ \text{N}_2 \]
Map a template pathway/network to the target genome

A naïve solution of pathway/network mapping

Orthology mapping through pair-wise sequence similarity comparison, e.g., bi-directional best hits (BDBH) method

Unfortunately, BDBH method does not always work!
Problems with the BDBH schema

Therefore, genes $g_1$ and $g_5$ are missing.
Mapping the P assimilation pathway of \textit{E. coli} to the WH 8102 genome

<table>
<thead>
<tr>
<th>E. coli template</th>
<th>Functions</th>
<th>Orthologs in WH8102</th>
</tr>
</thead>
<tbody>
<tr>
<td>pstB</td>
<td>ATP binding component</td>
<td>SYNW1272</td>
</tr>
<tr>
<td>pstA</td>
<td>intergal membrane protein</td>
<td>SYNW1271</td>
</tr>
<tr>
<td>pstC</td>
<td>intergal membrane protein</td>
<td>SYNW1270</td>
</tr>
<tr>
<td>pstS</td>
<td>P$_i$ binding protein</td>
<td>SYNW2507</td>
</tr>
<tr>
<td>phoB</td>
<td>response regulator</td>
<td>SYNW0947</td>
</tr>
<tr>
<td>phoR</td>
<td>sensor kinase</td>
<td>SYNW0948</td>
</tr>
<tr>
<td>phnE</td>
<td>channel protein</td>
<td>SYNW1168</td>
</tr>
<tr>
<td>phnD</td>
<td>periplasmic binding protein</td>
<td>SYNW1170</td>
</tr>
<tr>
<td>phnC</td>
<td>ATP binding component</td>
<td>SYNW1169</td>
</tr>
</tbody>
</table>

Information highly related to pathways/networks in prokaryotes

- **Operon structure**: In prokaryotes, proteins working together in a biological process are often encoded in an operon.

- **Regulon structure**: Several operons involved in the same pathway are often regulated by the same transcription factor, forming a regulon.

Pathway mapping
Using operon structure and *cis*-regulatory binding sites information

Define a pathway mapping graph:

\[ V_g : \text{template genes} \]
\[ V_o : \text{mapped operons} \]
\[ E_{go} \]

\[ g_1 \]
\[ g_2 \]
\[ g_3 \]
\[ g_4 \]
\[ g_5 \]
\[ g_m \]
\[ E_{oo} \]
\[ o_1 \]
\[ o_2 \]
\[ o_n \]
Pathway mapping
Formulate the pathway mapping problem as a constrained optimization problem
1. The overall sequence similarities between the homologous gene pairs are as high as possible
2. The involved operons share as many similar regulatory binding sites as possible.
3. The number of mapped operons is as small as possible

\[ V_g: \text{template genes} \]
\[ g_1 \rightarrow E_{go} \rightarrow o_1 \]
\[ g_2 \rightarrow E_{go} \rightarrow o_1 \]
\[ g_3 \rightarrow E_{go} \rightarrow o_1 \]
\[ g_4 \rightarrow E_{go} \rightarrow o_1 \]
\[ g_5 \rightarrow E_{go} \rightarrow o_1 \]
\[ g_6 \rightarrow E_{go} \rightarrow o_1 \]
\[ g_m \rightarrow E_{go} \rightarrow o_1 \]

\[ V_o: \text{mapped operons} \]
\[ o_1 \rightarrow E_{oo} \rightarrow o_2 \]
\[ o_1 \rightarrow E_{oo} \rightarrow o_2 \]
\[ o_1 \rightarrow E_{oo} \rightarrow o_n \]
\[ o_1 \rightarrow E_{oo} \rightarrow o_n \]
The goals of pathway mapping

1. The overall sequence similarities between the homologous gene pairs are as high as possible
2. The Number of mapped operons is as small as possible
3. The involved operons share as many similar regulatory binding sites as possible.
Pathway Mapping

---A solution by integer programming: PMAP program
http://csbl.bmb.uga.edu/pmap2/PMAP2.htm

Subject to:

\[
\sum_{i=1}^{n} \sum_{j=1}^{m} x_{ij} H_{ij} = \frac{\sum_{k=1}^{p} y_k S_k + \sum_{l=1}^{r} \sum_{k=1}^{p} z_{lk} B_{lk} + \sum_{l=1}^{r} u_l A_l}{m + n + p + r}
\]

Minimize objective function:

\[
\sum_{i=1}^{n} \sum_{j=1}^{m} x_{ij} H_{ij} + \sum_{k=1}^{p} y_k S_k + \sum_{l=1}^{r} \sum_{k=1}^{p} z_{lk} B_{lk} + \sum_{l=1}^{r} u_l A_l
\]

\[
\sum_{i=1}^{n} x_{ij} = 1, \quad \text{for } j=1..m \quad (1)
\]

\[
0 \leq \sum_{j=1}^{m} x_{ij} \leq 1, \quad \text{for } i=1..n \quad (2)
\]

\[
\sum_{i=1}^{n} \frac{\sum_{j=1}^{m} x_{ij}}{m \times n} \leq y_k \leq 1, \quad \text{for } k=1..p \quad (3)
\]

\[
\sum_{l=1}^{r} z_{lk} = y_k, \quad \text{for } k=1..p \quad (4)
\]

\[
\sum_{l=1}^{r} \frac{z_{lk}}{r \times p} \leq u_l \leq 1, \quad \text{for } l=1..r \quad (5)
\]

Mapping the template networks to the WH8102 genome using the PMAP program

From PCC7120 to WH8102

From PCC6803 to WH8102

From PCC7942 to WH8102

## Components in the initial network model

<table>
<thead>
<tr>
<th>Name</th>
<th>Protein ID</th>
<th>Synonym</th>
<th>Operon</th>
<th>Templates</th>
<th>SAM</th>
<th>Rank of NtcA binding site**</th>
</tr>
</thead>
<tbody>
<tr>
<td>amt1</td>
<td>33864789</td>
<td>synw0253</td>
<td></td>
<td>PCC6803</td>
<td></td>
<td></td>
</tr>
<tr>
<td>glnA</td>
<td>33865607</td>
<td>synw0173</td>
<td></td>
<td>PCC7120, PCC6803</td>
<td>-3.820</td>
<td>3</td>
</tr>
<tr>
<td>glnB</td>
<td>33864998</td>
<td>synw0462</td>
<td></td>
<td>PCC7120, PCC6803</td>
<td>6.422</td>
<td>26</td>
</tr>
<tr>
<td>gor</td>
<td>33866067</td>
<td>synw533</td>
<td></td>
<td>PCC7120</td>
<td></td>
<td></td>
</tr>
<tr>
<td>icd</td>
<td>33864702</td>
<td>synw0166</td>
<td></td>
<td>PCC7120, PCC6803</td>
<td>618</td>
<td></td>
</tr>
<tr>
<td>narB</td>
<td>33866994</td>
<td>synw2464</td>
<td>PCC7120, PCC6803, WH8103</td>
<td>-5.098</td>
<td>1</td>
<td></td>
</tr>
<tr>
<td>nirA</td>
<td>33867007</td>
<td>synw2477</td>
<td>PCC7120, PCC6803, WH8103</td>
<td>9</td>
<td></td>
<td></td>
</tr>
<tr>
<td>cynA</td>
<td>33867017</td>
<td>synw2487</td>
<td>synw2485 - synw2487</td>
<td>PCC7942</td>
<td>-5.132</td>
<td>4</td>
</tr>
<tr>
<td>cynB</td>
<td>33867016</td>
<td>synw2486</td>
<td>synw2485 - synw2487</td>
<td>PCC7942</td>
<td>-5.132</td>
<td>4</td>
</tr>
<tr>
<td>cynD</td>
<td>33867015</td>
<td>synw2485</td>
<td>synw2485 - synw2487</td>
<td>PCC7942</td>
<td>-5.132</td>
<td>4</td>
</tr>
<tr>
<td>cynS</td>
<td>33867020</td>
<td>synw2490</td>
<td></td>
<td>PCC7942</td>
<td>-3.420</td>
<td>735</td>
</tr>
<tr>
<td>nrtP</td>
<td>33866993</td>
<td>synw2463</td>
<td></td>
<td>WH8103</td>
<td>-5.184</td>
<td>960</td>
</tr>
<tr>
<td>ntcA</td>
<td>33864811</td>
<td>synw0275</td>
<td>synw0273 - synw0275</td>
<td>PCC7120, PCC6803</td>
<td>-5.098</td>
<td>1</td>
</tr>
<tr>
<td>petH</td>
<td>33865285</td>
<td>synw0751</td>
<td></td>
<td>PCC7120</td>
<td>-6.842</td>
<td>162</td>
</tr>
<tr>
<td>rbcL</td>
<td>33866250</td>
<td>synw1718</td>
<td></td>
<td>PCC7120</td>
<td></td>
<td></td>
</tr>
<tr>
<td>ureA</td>
<td>33866979</td>
<td>synw2449</td>
<td>synw2446 - synw2449</td>
<td>PCC7120</td>
<td>-3.602</td>
<td>98</td>
</tr>
<tr>
<td>ureB</td>
<td>33866978</td>
<td>synw2448</td>
<td>synw2446 - synw2449</td>
<td>PCC7120</td>
<td>-6.018</td>
<td>98</td>
</tr>
<tr>
<td>ureC</td>
<td>33866977</td>
<td>synw2447</td>
<td>synw2446 - synw2449</td>
<td>PCC7120</td>
<td>-7.659</td>
<td>98</td>
</tr>
<tr>
<td>ureD</td>
<td>33866976</td>
<td>synw2446</td>
<td>synw2446 - synw2449</td>
<td>PCC7120</td>
<td>98</td>
<td></td>
</tr>
<tr>
<td>ureE</td>
<td>33866975</td>
<td>synw2445</td>
<td>synw2443 - synw2445</td>
<td>PCC7120</td>
<td>78</td>
<td></td>
</tr>
<tr>
<td>ureF</td>
<td>33866974</td>
<td>synw2444</td>
<td>synw2443 - synw2445</td>
<td>PCC7120</td>
<td>78</td>
<td></td>
</tr>
<tr>
<td>ureG</td>
<td>33866973</td>
<td>synw2443</td>
<td>synw2443 - synw2445</td>
<td>PCC7120</td>
<td>78</td>
<td></td>
</tr>
<tr>
<td>urtA</td>
<td>33866972</td>
<td>synw2442</td>
<td>synw2438 - synw2442</td>
<td>PCC7120</td>
<td>2</td>
<td></td>
</tr>
<tr>
<td>urtB</td>
<td>33866971</td>
<td>synw2441</td>
<td>synw2438 - synw2442</td>
<td>PCC7120</td>
<td>-4.551</td>
<td>2</td>
</tr>
<tr>
<td>urtC</td>
<td>33866970</td>
<td>synw2440</td>
<td>synw2438 - synw2442</td>
<td>PCC7120</td>
<td>-2.995</td>
<td>2</td>
</tr>
<tr>
<td>urtD</td>
<td>33866969</td>
<td>synw2439</td>
<td>synw2438 - synw2442</td>
<td>PCC7120</td>
<td>-4.429</td>
<td>2</td>
</tr>
<tr>
<td>urtE</td>
<td>33866968</td>
<td>synw2438</td>
<td>synw2438 - synw2442</td>
<td>PCC7120</td>
<td>-4.423</td>
<td>2</td>
</tr>
</tbody>
</table>

Initial model of nitrogen assimilation network in WH8102
Expansion of the initial network model

Rationale

• The initial model is likely incomplete
• Filling gaps through recruiting additional genes into the pathway model

Methods----Guilty by association

• Prediction of operon structures
• Prediction of protein-protein interaction map
• Prediction of functional association between proteins through phylogenetic profile analyses
• Prediction of *cis*-regulatory binding sites for the response regulators in the pathways/networks
Recruiting genes through prediction of protein-protein interactions

Homology mapping

DIP database

Target genome

Detection of gene fusion events

Target genome

nr Database

$\text{Score} = -\frac{1}{2} (\log E_1 + \log E_2)$
Validation on *E. coli* data set in DIP

\[
p(K_V | s_c) = 1 - \sum_{i=1}^{K_V} \binom{K_E}{i} \frac{(N-K_E)}{N} \binom{K_P}{i} \frac{K_P}{N-K_E}
\]

\(N=9,294,516;\) and \(K_E = 457;\)

Prediction by homology mapping

\(K_P=831;\) \(K_V=31;\) \(P<10^{-30}\)

Prediction by protein fusion analyses

\(K_P=1807;\) \(K_V=17;\) \(P<10^{-8}\)

Therefore, predictions with high statistical significance can be achieved by both methods

Predicted protein interaction map in WH8102

Recruiting genes through phylogenetic profile analysis

2-D representation of the clustering results

Recruiting Genes through phylogenetic profile analysis

\[ d_i \]

\[ P < 10^{-10} \]

\[ d_i \]

\[ P < 0.0448 \]

\[ P < 0.0175 \]

\[ P < 0.057 \]

Mapping coordinates

Mapping coordinates
Phylogenetic footprinting

Orthologs identification

\[ G_i.g_s \rightarrow G_j.g_t \rightarrow \ldots \]

\[ G_1.g_s \rightarrow G_1.g_t \rightarrow \ldots \]

\[ G_2.g_s \rightarrow G_2.g_t \rightarrow \ldots \]

\[ G_n.g_s \rightarrow G_n.g_t \rightarrow \ldots \]

Intergenic regions

Motif finding
Using CUBIC

Scan the intergenic regions of the genome

Putative binding sites

PSWM

CUBIC: http://csbl.bmb.uga.edu/downloads/#cubic
High conservation of the DNA binding domain of NtcA protein

Biological bases of the algorithm

3. *cis*-binding sites have a higher probability to occur in the intergenic regions than in the coding regions. Thus, a statistical model can be designed to distinguish the true binding sites from the spurious ones occurring by chance.

\[ LOR(s) = \ln \frac{p(S(I) > s)}{p(S(C) > s)} \]

Phylogenetic footprinting

Two highly conserved motifs are identified in the regulatory regions of genes known being regulated by NtcA across 9 cyanobacterial genomes.

NtcA binding sites

$\sigma^{70}$-factor binding sites
Predicted NtcA promoters in WH8102 at p<0.05

<table>
<thead>
<tr>
<th>Rank</th>
<th>Synonym</th>
<th>Name</th>
<th>NtcA site</th>
<th>Downstream of NtcA binding sites</th>
<th>NtcA site position</th>
<th>Score</th>
<th>-10k box Score</th>
</tr>
</thead>
</table>
Validation of the predicted components in the network

Culture WH8102

1mM NH₄⁺  9mM Na₂NO₃

Cy3  Cy5

mRNA  Reverse transcription

cDNA

mix samples

Hybridize

Glass slide microarray

Analysis with SAM package

SAM score=

\[ \begin{align*}
&\leq -\alpha, - \\
&> \alpha, +
\end{align*} \]
Validation of the recruited components in the network

- 247 genes were down-regulated by NH$_4^+$, 8 of which are photosynthetic genes.
- 91 genes were up-regulated by NH$_4^+$, 8 of which are photosynthetic genes.
- P-value:

\[
 p(K,V) = 1 - \sum_{i=1}^{V} \frac{A\binom{n-A}{i}\binom{n}{K-i}}{n^K} , \quad n = 2516 \text{ genes}
\]

<table>
<thead>
<tr>
<th>Methods</th>
<th>Number of genes recruited</th>
<th>Number of genes affected</th>
<th>p-value &lt;</th>
</tr>
</thead>
<tbody>
<tr>
<td>Initial model</td>
<td>27</td>
<td>14</td>
<td>1.42X10^{-6}</td>
</tr>
<tr>
<td>Protein-protein interactions</td>
<td>11</td>
<td>4</td>
<td>1.33x10^{-6}</td>
</tr>
<tr>
<td>Phylogenetic profile</td>
<td>5</td>
<td>3</td>
<td>2.93x10^{-4}</td>
</tr>
<tr>
<td>Regulon prediction</td>
<td>89</td>
<td>64</td>
<td>1.55X10^{-24}</td>
</tr>
<tr>
<td>Operon</td>
<td>2</td>
<td>1</td>
<td>0.018</td>
</tr>
<tr>
<td>Combined</td>
<td>124</td>
<td>74</td>
<td>5.6X10^{-40}</td>
</tr>
</tbody>
</table>

Validation by whole genome microarray gene expression experiments

\[ \text{Score} \]

\[ \text{Probability or LOR} \]

- \( p(S(g_d > s)) \)
- \( p(S(g_o > s)) \)
- \( LOR \)

Graph showing the relationship between score and probability or LOR.
A working model for the N assimilation network in WH8102

Computational Protocol for Microbial Pathway Inference

I. Literature/database search
   - Template pathways in relevant species
     - Genomic sequences
       - Orthologues
       - Operon structures
         - Sub-pathways
           - Initial pathways
             - Expansion and Refinement
               - Predicted pathways
                 - Validation
               - Fusion events
                 - Protein-protein interaction map
                   - Yeast two hybrid data

II. Expansion and Refinement
III. Validation
IV. Fusion events

- Phylogenetic profiles
- Regulatory binding sites
Conclusions from the working model

1. The global nitrogen assimilation network in WH8102 involves probably more than 400 genes;

2. These genes are under control of a transcription regulatory network consisting of global regulators (NtcA, SigA and RpoD) and 8 putative specific regulators;

3. Nitrogen assimilation and photosynthesis are two tightly coupled biological processes, and we postulate that photosynthetic genes that bear NtcA promoters are the bridging points between the two biological processes;

4. The model should be very useful for rational design of further detailed experimental investigations.