Machine Learning pathway models of polysaccharide structures from Campylobacter jejuni

Stephen Muggleton
Centre for Integrative Systems Biology
Imperial College, London
April, 2010
Overview

CISBIC
Machine Learning
CISBIC sub-project 1
Hypotheses being tested
Automated experiment selection
Conclusions
Systems Biology: The CISBIC Vision
Exemplar Research Project

• **Host-pathogen interactions**
  • combating infectious disease in plants and animals
  • using infection as a perturbation to study host biology
<table>
<thead>
<tr>
<th>Logical</th>
<th>Probabilistic</th>
<th>Mixed</th>
</tr>
</thead>
<tbody>
<tr>
<td>Decision trees</td>
<td>Neural nets</td>
<td>Bayes’ nets</td>
</tr>
<tr>
<td>Grammars</td>
<td>HMMs</td>
<td>SCFGs</td>
</tr>
<tr>
<td>Logic Programs</td>
<td>POMDPs</td>
<td>BLPs</td>
</tr>
</tbody>
</table>
Inductive Logic Programming

**Background knowledge.** Protein sequence, partial grammar, incomplete biological network.

**Examples.** Molecules, annotated sentences, temporal traces of up/down regulation.

**Hypothesis.** Explanation of molecular 3-D shape, new clauses in a grammar, extra network annotation.
### ILP for Systems Biology

<table>
<thead>
<tr>
<th>Biology</th>
<th>Robot Scientist</th>
<th>Metalog</th>
<th>CISBIC</th>
</tr>
</thead>
<tbody>
<tr>
<td>Computing</td>
<td>Active learning</td>
<td>Networks</td>
<td>Dynamic Modelling</td>
</tr>
<tr>
<td>ETAI 2001</td>
<td></td>
<td>MLJ 2006</td>
<td></td>
</tr>
</tbody>
</table>
**Aim.** Model effect of changes to pathogen genome on expression of glycosylated surface molecules involved in triggering innate immune response of host.
Initial focus

• Focussing initially on synthetic pathways related to construction of capsule structure for Campylobacter jejuni.
• 38 genes involved. Fully sequenced. Have full set of knock-outs.
• Bile-salt perturbation microarray and metabonomic experiments.
• 1/3 functions already known, 1/3 suspected and 1/3 unknown.
Pathway database

Extraction / Visualisation of Campylobacter pathways

KEGG
Campylobacter
E Coli
Helicobacter

BioCyc
Campylobacter
E Coli
Helicobacter

Brendan/papers
Campylobacter

Prolog database

Machine learning

Visualisation
Examples of Prolog representation

reaction(kegg,cje,r04606,c04932,c04652,2).

reaction_gene(kegg,cje,r04657,cj0811).

reaction_EC(kegg,cje,r04657, "2.7.1.130").

compound_name(kegg,c04932,"2,3,2',3'-Tetrakis(3-hydroxytetradecanoyl)-D-glucosaminyl-1,6-beta- D-glucosamine 1-phosphate").

xlink_compound(biocyc,bischomyrglc,kegg,c04932).

gene_name(kegg,cj0811,"tetraacyldisaccharide 4'-kinase").

reaction_pathway(kegg,cje,r04606,path00540).

pathway_name(kegg,cje,path00540,"Lipopoly saccharide biosynthesis").
An Abductive framework for modeling genotype-phenotype relation

observable(g1,s1).
up(g1,c1).
down(g1,c3).

observable(G,S):-
  struct_next(S,S2),
  start(S),
  path(S,X),
  reaction(R,X,Y),
  codes(G,R),
  path(Y,S2),

codes(g1,r1).

topology:
  reaction(r1,c1,c2).
  reaction(r2,c3,c4).
  struct_next(s1,s2)
  struct_next(s3,s4)

Pathways

Glycans

New Observations

Experimental Evaluations (in vivo)

Hypotheses

New Theory
Pathway abduction

The current pathways can only explain around 34% of mutant data.
Example of hypotheses for capsule pathway (in red)
Hypotheses predict

1. Cj1416, Cj1417, Cj1418 involved in OMePN synthesis. Wren group has verified this by mutagenesis and structural analysis. Fairly obvious from BLAST analysis.

2. Cj1432, a protein of unknown function, is central to capsule synthesis. Not predictable from amino acid similarity and BLAST analysis.
Ongoing experiments to test cj1432 hypothesis

1. Immunoassay - initial results suggest absence of capsule
2. Alcian Blue dye staining of Campylobacter initial results suggest absence of capsule
3. Electron microscopy - direct visual inspection
4. Complementation - reintroduce cj1432 to chromosome of mutant
5. Structural analysis of capsule glycan - NMR, MS, lipid-linked intermediates
6. Protein purification and enzyme assays
### Hypotheses / Experiments decision trees

![Decision Trees Diagram](image)

<table>
<thead>
<tr>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td>H1: codes(geneX,rA)</td>
<td>0</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>0</td>
<td>1</td>
<td>1</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>H2: codes(geneX,rB)</td>
<td>1</td>
<td>0</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>1</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>H3: codes(geneX,rC)</td>
<td>1</td>
<td>1</td>
<td>0</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>0</td>
<td>1</td>
<td>1</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>H4: codes(geneX,rD)</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>0</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>0</td>
<td>1</td>
<td>0</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>H5: codes(geneX,rE)</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>0</td>
<td>1</td>
<td>1</td>
<td>0</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>0</td>
</tr>
</tbody>
</table>
Conclusions and further work

- Integration of diverse background knowledge
- ILP produces readable rules
- Adbuction for gap-filling in networks
- Gap-filling vs island hopping
- Ongoing NMR/MS experimental testing of hypotheses
- cj4132 - potential C. jejeuni vaccine?