Inferring protein-protein interactions from amino acid sequences:
Application to two-component systems

Most of this can be found in:
A Bayesian algorithm for reconstructing two-component signaling networks
Lukas Burger and Erik van Nimwegen
Proceedings 6th International Workshop, WABI 2006, Zurich, Switzerland

Lukas Burger and Erik van Nimwegen

Division of Bioinformatics
Biozentrum, Universität Basel,
Swiss Institute of Bioinformatics
Bacterial signaling: Two-component systems

- Responsible for most signal transduction in bacteria.
  - E. coli: aerobic/anaerobic switch, chemotaxis
  - B. subtilis: sporulation
  - C. crescentus: cell-cycle/differentiation.
- Over 8,500 two-component system genes in 399 sequenced bacterial genomes.
Advantages as model system for predicting protein-protein interactions

- Two-component systems can be easily detected using hidden Markov models of the kinase and receiver domains.
- Enough homology to reproduce reliable multiple alignments: specificity of interaction likely lies in details of amino acids at surface.
- Large number of examples available (good statistics).
- *Training sets*: For about 50% we *know* which kinase interacts with receiver because they lie in a common operon.

\[ \text{cognate pairs} \quad \text{orphan kinases and regulators} \]
Extracting two-component system proteins

- **Receivers**: All HMMer hits to Pfam profile `Response_reg`.
- **Kinases**: All HMMer hits to the Pfam profiles:
  - HisKA, H2, H3, His_kinase, HWE_HK, HATPase_c, HPT

<table>
<thead>
<tr>
<th>Name</th>
<th>Architecture</th>
<th>no.cognates</th>
<th>no.orphans</th>
</tr>
</thead>
<tbody>
<tr>
<td>HisKA</td>
<td>HisKA, HATPase_c</td>
<td>3388</td>
<td>2158</td>
</tr>
<tr>
<td>H3</td>
<td>H3, HATPase_c</td>
<td>636</td>
<td>183</td>
</tr>
<tr>
<td>His_kinase</td>
<td>HATPase_c</td>
<td>245</td>
<td>23</td>
</tr>
<tr>
<td>Long hybrid</td>
<td>HisKA, HATPase_c, RR, (RR), Hpt</td>
<td>132</td>
<td>286</td>
</tr>
<tr>
<td>Short hybrid</td>
<td>HisKA, HATPase_c, RR, (RR)</td>
<td>126</td>
<td>985</td>
</tr>
<tr>
<td>Chemotaxis</td>
<td>Hpt, HATPase_c</td>
<td>89</td>
<td>77</td>
</tr>
<tr>
<td>Hpt</td>
<td>Hpt</td>
<td>37</td>
<td>192</td>
</tr>
<tr>
<td>HWE</td>
<td>HWE or H2, HATPase_c</td>
<td>34</td>
<td>162</td>
</tr>
</tbody>
</table>

Table 1: Pfam domain combinations of the most abundant kinase architectures and the numbers of their occurrence in both cognates and orphans. Both the short and long hybrid architecture can contain one or two receiver domains.
Global statistics `signaling networks’ in bacteria

Total number of two-comp. proteins

Fraction of all two-comp. that are orphans

Total number of orphans

Fraction of receivers among orphans
Multiple Alignments

- Aligned all receivers together. Separate alignment for each kinase domain architecture.
- Produced multiple alignments by aligning to the Pfam profiles.
- Produced independent alignments with ProbCons.
- Keep only positions with greater than 80% agreement by the two methods and less than 50% gaps.
- Gives alignments with absolute reference positions for all kinase and receiver sequences.

**Final number of positions used:**
Receiver domain: 115.
Classifying receivers

- For each class $c$, each position $i$ and each amino acid $\alpha$ estimate the weight matrix of probabilities to obtain letter $\alpha$ at position $i$ of a receiver in class $c$ from the cognate kinase/receiver pairs.

**Weight matrix:**

$$w_{ci}^c = \frac{n_{ci}^c + \lambda}{n_c^c + 21\lambda}$$

**Probability receiver given class:**

$$P(\vec{R} \mid c) = \prod_i w_{R_i}^c$$

**Posterior class given receiver:**

$$P(c \mid \vec{R}) = \frac{P(\vec{R} \mid c)P(c)}{\sum_{\tilde{c}} P(\vec{R} \mid \tilde{c})P(\tilde{c})}$$

**Some technical details:**

- To correct for phylogenetic sampling biases we single-link cognates into clusters of similarity $\geq 90\%$ and give a weight of 1 to each cluster.
- $\lambda$ is set to $\frac{1}{2}$ (Prior uniform in Fisher-information).
- Gaps are treated as 21st amino acid.
- To get cognates we consider genes separated by $\leq 50$ bp on same strand in same operon and take only cognates if there is only 1 kinase and 1 receiver in the operon.
- We use a uniform prior $P(c)$.
- We take out the receiver and all members of its cluster from the counts in the WM when scoring a given receiver.
Results Classifying receivers

Overall 94% of cognate receivers are classified correctly
Interacting amino acids

\( \vec{k} = (k_1, k_2, \cdots, k_n) \)  Amino acid sequence of a kinase.

\( \vec{r} = (r_1, r_2, \cdots, r_m) \)  Amino acid sequence of a receiver.

\( (\vec{k}, \vec{r}) = \vec{s} \)  Joint sequence of interacting kinase/receiver pair.

\[ P(\vec{k}, \vec{r}) \equiv P(\vec{s}) \]  Joint probability of observing the joint sequence for an interacting kinase/receiver pair.

We will factor joint distribution into conditional probabilities of pairs of positions.

\[ P(\vec{s}) = P(s_r) \prod_{i \neq r} P(s_i | s_{\pi(i)}) \]

\( s_r \) = Amino acid at root.

\( \pi(i) \) = position that amino acid at position \( i \) depends on.

Example:
Probability of the data given a dependence tree topology

\[ P(s_i \mid s_j) = \rho_{s_is_j}^{ij} \]  

Parametrization of the conditional probabilities.

\[ n_{\alpha\beta}^{ij} \]  

Number of times amino acid combination \( \alpha, \beta \) occurs at positions \( i,j \) in the data.

\[
P(D \mid \rho, \pi) = \left[ \prod_{\alpha} \left( \rho_{\alpha}^r \right)^{n_{\alpha}^r} \right] \prod_{i} \left[ \prod_{\alpha\beta} \left( \rho_{\alpha\beta}^{i\pi(i)} \right)^{n_{\alpha\beta}^{i\pi(i)}} \right]
\]  

Likelihood

\[
P(\rho \mid \pi) \propto \left[ \prod_{\alpha} \left( \rho_{\alpha}^r \right)^{21\lambda - 1} \right] \prod_{i} \left[ \prod_{\alpha\beta} \left( \rho_{\alpha\beta}^{i\pi(i)} \right)^{\lambda - 1} \right]
\]  

Prior

\[
P(D \mid \pi) = \int P(D \mid \rho, \pi) P(\rho \mid \pi) d\rho
\]  

Likelihood of the dependence tree
Probability given dependence tree topology

\[ P(D_i \mid \pi, D_j) = \prod_{\beta} \left[ \frac{\Gamma(21\lambda) \int \prod_{\alpha} \frac{(\rho_{\alpha\beta}^{ij})^{n_{ij}^{\alpha\beta}}}{\Gamma(\lambda)} d\rho_{\alpha\beta}^{ij}}{\Gamma(21\lambda)} \right] = \right. \\
\left. \prod_{\beta} \left[ \frac{\Gamma(21\lambda)}{\Gamma(n_{\beta}^j + 21\lambda)} \right] \prod_{\alpha} \left[ \frac{\Gamma(n_{\alpha\beta}^{ij} + \lambda)}{\Gamma(\lambda)} \right] \equiv M_i R_{ij} \right. \\

 Marginal and edge probabilities:

\[ M_i = \prod_{\alpha} \frac{\Gamma(n_{\alpha}^i + 21\lambda)}{\Gamma(21\lambda)} \]
\[ R_{ij} = \prod_{\alpha\beta} \frac{\Gamma(n_{\alpha\beta}^{ij} + \lambda) \Gamma(21\lambda) \Gamma(21\lambda)}{\Gamma(\lambda) \Gamma(n_{\alpha}^i + 21\lambda) \Gamma(n_{\beta}^j + 21\lambda)} \]

Final expression:

\[ P(D \mid \pi) = \frac{\Gamma(21^2 \lambda)}{\Gamma(n + 21^2 \lambda)} \left[ \prod_i M_i \right] \left[ \prod_{i \neq r} R_{i\pi(i)} \right] \]
Maximal Likelihood dependence tree

\[
P(D \mid \pi) = \frac{\Gamma(21^2 \lambda)}{\Gamma(n + 21^2 \lambda)} \left[ \prod_i M_i \right] \left[ \prod_{i \neq r} R_{i \pi(i)} \right]
\]

\( \pi_* \) Dependence tree that maximizes \( P(D \mid \pi) \)

**Chow-Liu algorithm**: Start with a complete graph with weights \( R_{ij} \) on edges \((i,j)\). One can find the *maximal spanning tree* of this graph by a simple greedy procedure.

**Test: predicting interaction partners for cognate pairs.**

- Use dataset \( D \) of all cognate pairs to determine best dependence tree.
- For each genome, remove all its cognate pairs from the data-set, and remove also all single-linkage clusters of cognates with \( \geq 90\% \) amino acid identity.
- `Freeze' all other cognate kinase/receiver pairs (training set).
- Use Monte-Carlo to search over all *assignments* of kinase/receiver pairs for the genome under study.
- Probability of assignment is probability of joint data (frozen pairs + assignment).
Predicting Cognate interacting pairs

Genome being sampled:

\[ D_a \]

Likelihood assignment:

\[ P(D_f, D_a | \pi_*, \alpha) \]

Monte-Carlo sampling according to likelihood:

At every time-point pick 2 kinases at random and consider flipping the receivers assigned to them.
Predicting Cognate interacting pairs

Fraction of predictions that are true cognate pairs.

Remove all with >90% homology from training set.
Remove all with >80% homology from training set.
Phylogeny signal vs. physical interactions

- Do positions in kinase receiver just look correlated because there are orthologous interacting pairs are evolutionarily related?
- Make a new data-set of ‘false interacting pairs’ that have the exact same evolutionary relationships: Flip assignment of pairs for orthologous groups.

**True interacting pairs**

- Genome 1
  - K1
  - K2
  - R1
  - R2

- Genome 2
  - K1'
  - K2'
  - R1'
  - R2'

- Genome 3
  - K1''
  - K2''
  - R1''
  - R2''

**False interacting pairs**

- Genome 1
  - K1
  - K2
  - R2
  - R1

- Genome 2
  - K1'
  - K2'
  - R2'
  - R1'

- Genome 3
  - K1''
  - K2''
  - R2''
  - R1'
Number of pairs of positions with `interaction score’ $\log(R_{ij})$ over a given cut-off in true and false pairs.
Predicting orphan interactions

Extensions for doing orphan predictions:

• Use entire set of cognates as a `frozen’ training set.

• Run assignments of all orphans from all genomes at the same time.

• Run on all classes of kinases at the same time.

• Since the number of kinases and receivers are not the same, some kinases or receivers will not be hooked up to any partner.

• Each receiver can belong to each of the kinase classes (sampled over).

• Unhooked kinases/receivers are scored according to the simple WM model: one WM for each class of receiver and one for each class of kinase.

• Move-set includes changes in receiver class, and flips of partner (where one of the two members may not have a partner).
Results predicting orphan interactions

Focused on Caulobacter crescentus for which most results are available. C. crescentus has 11 orphan HisKA kinases and 19 orphan receivers.

<table>
<thead>
<tr>
<th>kinase</th>
<th>regulator</th>
<th>posterior</th>
<th>std</th>
<th>exp evidence</th>
</tr>
</thead>
<tbody>
<tr>
<td>CC0248</td>
<td>CC0247</td>
<td>1.0000</td>
<td>0.0000</td>
<td>putative cognate pair</td>
</tr>
<tr>
<td>CC0250</td>
<td>CC0294</td>
<td>0.0931</td>
<td>0.0078</td>
<td>cognate pair, in vitro phosphorylation [13]</td>
</tr>
<tr>
<td>CC2765</td>
<td>CC2766</td>
<td>0.5902</td>
<td>0.0154</td>
<td>cognate pair, in vitro phosphorylation [13]</td>
</tr>
<tr>
<td>CC2932</td>
<td>CC2931</td>
<td>0.9759</td>
<td>0.0213</td>
<td>putative cognate pair</td>
</tr>
<tr>
<td>CC2755</td>
<td>CC2757</td>
<td>0.8881</td>
<td>0.1613</td>
<td>putative cognate pair</td>
</tr>
<tr>
<td>CenK</td>
<td>CenR</td>
<td>0.7980</td>
<td>0.1601</td>
<td>in vitro phosphorylation [13]</td>
</tr>
<tr>
<td>ChpT</td>
<td>CpdR</td>
<td>0.7766</td>
<td>0.1453</td>
<td>in vitro phosphorylation [5]</td>
</tr>
<tr>
<td>pleC</td>
<td>DivK</td>
<td>0.6639</td>
<td>0.2088</td>
<td>in vitro phosphorylation [13]</td>
</tr>
<tr>
<td>CckN</td>
<td>CtrA</td>
<td>0.6379</td>
<td>0.1606</td>
<td>not known</td>
</tr>
<tr>
<td>DivL</td>
<td>CC0588</td>
<td>0.6187</td>
<td>0.2272</td>
<td>not known</td>
</tr>
<tr>
<td>DivJ</td>
<td>PleD</td>
<td>0.4890</td>
<td>0.2190</td>
<td>in vitro phosphorylation [13]</td>
</tr>
<tr>
<td>CckN</td>
<td>PleD</td>
<td>0.2934</td>
<td>0.1143</td>
<td>not known</td>
</tr>
<tr>
<td>DivJ</td>
<td>CtrA</td>
<td>0.2147</td>
<td>0.1269</td>
<td>in vitro phosphorylation [18]</td>
</tr>
<tr>
<td>PleC</td>
<td>PleD</td>
<td>0.1650</td>
<td>0.1332</td>
<td>in vitro phosphorylation [13]</td>
</tr>
<tr>
<td>CenK</td>
<td>CC0588</td>
<td>0.1260</td>
<td>0.0913</td>
<td>false positive, in vitro phosphorylation [13]</td>
</tr>
<tr>
<td>DivJ</td>
<td>DivK</td>
<td>0.1045</td>
<td>0.0401</td>
<td>in vitro phosphorylation [13]</td>
</tr>
<tr>
<td>DivL</td>
<td>DivK</td>
<td>0.0964</td>
<td>0.0693</td>
<td>yeast two-hybrid screen [10]</td>
</tr>
<tr>
<td>CC2755</td>
<td>CC0588</td>
<td>0.0948</td>
<td>0.1424</td>
<td>not known</td>
</tr>
<tr>
<td>DivJ</td>
<td>CenR</td>
<td>0.0881</td>
<td>0.0615</td>
<td>false positive, in vitro phosphorylation [13]</td>
</tr>
<tr>
<td>ChpT</td>
<td>DivK</td>
<td>0.0877</td>
<td>0.1256</td>
<td>false positive, in vitro phosphorylation [5]</td>
</tr>
<tr>
<td>DivJ</td>
<td>CC0588</td>
<td>0.0638</td>
<td>0.0214</td>
<td>false positive, in vitro phosphorylation [5]</td>
</tr>
<tr>
<td>DivL</td>
<td>CtrA</td>
<td>0.0637</td>
<td>0.0357</td>
<td>in vitro phosphorylation [18]</td>
</tr>
<tr>
<td>CenK</td>
<td>CC0432</td>
<td>0.0608</td>
<td>0.1216</td>
<td>not known</td>
</tr>
<tr>
<td>DivL</td>
<td>PleD</td>
<td>0.0574</td>
<td>0.0600</td>
<td>not known</td>
</tr>
<tr>
<td>DivL</td>
<td>CC0432</td>
<td>0.0479</td>
<td>0.0958</td>
<td>not known</td>
</tr>
<tr>
<td>ChpT</td>
<td>CtrA</td>
<td>0.0415</td>
<td>0.0324</td>
<td>in vitro phosphorylation [5]</td>
</tr>
<tr>
<td>CckN</td>
<td>DivK</td>
<td>0.0407</td>
<td>0.0345</td>
<td>yeast two-hybrid screen [10]</td>
</tr>
</tbody>
</table>

Top predictions for hisKA kinases with known interactions. (of 319 possible interactions). List was cut to have at least 1 prediction for all kinases for which an interaction is known.

- **Green**: Experimentally confirmed
- **Blue**: Interaction observed in yeast two-hybrid screen.
- **Yellow**: No data available either for or against.
- **Red**: Interaction not observed experimentally when tested.

Probability to get this match by chance: 

\[ p = 1.4 \times 10^{-13} \]
Results predicting orphan interactions

One hisKA orphan interaction known in Helicobacter pylori.

<table>
<thead>
<tr>
<th>kinase</th>
<th>regulator</th>
<th>posterior</th>
<th>std</th>
<th>exp evidence</th>
</tr>
</thead>
<tbody>
<tr>
<td>HP0244</td>
<td>HP0703</td>
<td>1</td>
<td>0</td>
<td>in vitro phosphorylation [4]</td>
</tr>
<tr>
<td>HP0244</td>
<td>HP1067</td>
<td>0</td>
<td>0</td>
<td></td>
</tr>
<tr>
<td>HP0244</td>
<td>HP1043</td>
<td>0</td>
<td>0</td>
<td></td>
</tr>
<tr>
<td>HP0244</td>
<td>HP1021</td>
<td>0</td>
<td>0</td>
<td></td>
</tr>
<tr>
<td>HP0244</td>
<td>HP0616</td>
<td>0</td>
<td>0</td>
<td></td>
</tr>
<tr>
<td>HP0244</td>
<td>HP0393</td>
<td>0</td>
<td>0</td>
<td></td>
</tr>
<tr>
<td>HP0244</td>
<td>HP0019</td>
<td>0</td>
<td>0</td>
<td></td>
</tr>
</tbody>
</table>

Known interaction is predicted with posterior 1.
There are 7 orphan receivers in Helicobacter pylori.
Beyond maximum-likelihood spanning tree: summing over spanning trees

Probability of the data given a spanning dependence tree:

\[
P(D \mid \pi) = \frac{\Gamma(21^2 \lambda)}{\Gamma(n + 21^2 \lambda)} \left[ \prod_i M_i \right] \left[ \prod_{i \neq r} R_{i \pi(i)} \right]
\]

We would like to sum over all possible trees:

\[
P(D) = \sum_{\pi} \frac{P(D \mid \pi)}{|\pi|} = \frac{\Gamma(21^2 \lambda)}{|\pi| \Gamma(n + 21^2 \lambda)} \left[ \prod_i M_i \right] \sum_{\pi} \left[ \prod_{i \neq r} R_{i \pi(i)} \right]
\]

Example: for 3 positions we would sum over the three spanning trees:

\[
P(D) \propto R_{12} R_{13} + R_{13} R_{23} + R_{12} R_{23}
\]
Generalization of the matrix-tree theorem

Define Laplacian matrix:
\[ L_{ij} = \delta_{ij} \sum_k R_{ik} - R_{ij} \]

\[ \tilde{L} = \text{Matrix } L \text{ with a single row and column removed} \]

Theorem:
\[ \sum_{\pi} \left[ \prod_i R_{i\pi(i)} \right] = \det(\tilde{L}) \]

One catch: \( R \) has some components that are extremely large and others that are extremely small. We so far have found no numerically stable way of calculating the determinant, only uncontrolled approximations.

Test:
- Take all cognate kinases and receivers and run Monte-Carlo assigning all genomes at the same time.
- Score the combination of assignments from all genomes using the determinant.
- Note: No training set!
Caveat: We cannot show that the Monte-Carlo has converged (and believe it has in fact not).