SIMULTANEOUS INERENCE OF BIOLOGICAL NETWORKS OF MULTIPLE SPECIES FROM GENOME-WIDE DATA AND EVOLUTIONARY INFORMATION: A SEMI-SUPERVISED APPROACH

Hisashi Kashima, Yoshihiro Yamanishi, Tsuyoshi Kato, Masashi Sugiyama, Koji Tsuda

Bioinformatics 2009 (Vol. 25 Issue 22)
Link Propagation - new semi-supervised learning method for inferring biological networks of multiple species

Motivation: use both cross-species and intra-species knowledge to infer biological networks
DATA FOR INFERRING BIOLOGICAL NETWORKS

- **Intra-species:**
  - gene order information, phylogenetic profiles, *gene expression* patterns, ...

- **Cross-species:**
  - evolutionary information about the conservation of protein interactions (interlog)
METHODS FOR INFERRING BIOLOGICAL NETWORKS

- Metric learning
  - Kernel canonical correlation analysis, dimension reduction, em-algorithm

- Binary classification
  - E.g. SVM with pairwise kernels (P-SVM)

- Applicable to various networks
- Algorithm needs optimization
**Link Propagation**

- Based on *label propagation* method 
  *(Zhou et al., 2004; Zhu et al., 2003)*
- Semi-supervised learning algorithm
- Uses both intra-species and cross-species information
- Simultaneous inference of multiple networks
  - (inferring metabolic networks of 3 species)
- Computationally efficient
  - P-SVM – time $O(m^6)$, space $O(m^4)$
  - Link prop. - time $O(m^5)$, space $O(m^2)$
LINK PROPAGATION

The principle of link propagation when there is a link

The principle of link propagation when no link exists
**Materials**

- **Species:** *Caenorhabditis elegans, Helicobacter pylori, Saccharomyces cerevisiae*
- **Gold standard (metabolic network)**
  - *KEGG PATHWAY DB*
  - **nodes/edges:** 532/2392, 291/492, 722/2323
- **Cross-species sequence similarities**
  - *KEGG GENES DB*
- **Gene expression data**
  - *MSGR DB*
CROSS-SPECIES SEQUENCE SIMILARITIES

- **KEGG GENES** - amino acid sequences
- Similarity between 2 proteins:

\[
s(g, g')/(\sqrt{s(g, g)} \sqrt{s(g', g')})
\]

\(s(x, y)\) - Smith–Waterman score
**GENE EXPRESSION DATA**

- **MSGR** (multiple-species gene recommender)
  - For each species:
    - 1209, 293, 753 microarray hybridization measurements
  - For each protein – a vector \((x)\) of that dimension
  - Similarity: Gaussian RBF kernel

\[
k(x, x') \equiv \exp\left(-\frac{||x - x'||^2}{2\gamma^2}\right) \text{ with } \gamma \equiv 2.
\]
METHOD - GOAL

Input:
- Adjacency matrices
  - Elements: \{-1,0,1\}
- Similarity matrices
  - Elements: Gene expression & Cross-species sequence similarities

Output:
- Link strength matrices
  - Elements: link strength – how likely is it that a link exists
**Method - Mathematics**

- If two pairs of nodes are similar to each other, then the two pairs have similar link strengths

\[
J(F) \equiv \sigma \frac{\text{vec}(F)^\top L \text{vec}(F)}{2} + \frac{1}{2} \| \text{vec}(F) - \text{vec}(A^*) \|_2^2
\]

\[
[A^{(k)*}]_{i(k), j(k)} \equiv \begin{cases} 
\frac{|A^{(k)}_{j(k)}| + |A^{(k)}_{i(k)}|}{|A^{(k)}_{j(k)}| + |A^{(k)}_{i(k)}|} & \text{if } [A^{(k)}]_{i(k), j(k)} = 1, \\
-\frac{|A^{(k)}_{j(k)}| + |A^{(k)}_{i(k)}|}{|A^{(k)}_{j(k)}| + |A^{(k)}_{i(k)}|} & \text{if } [A^{(k)}]_{i(k), j(k)} = -1, \\
0 & \text{otherwise.}
\end{cases}
\]

\[
L \equiv \begin{bmatrix} \tilde{D}^{(1)} & 0 \\
\vdots & \ddots & 0 \\
0 & \ddots & \ddots & \ddots \\
& \ddots & \ddots & \ddots & \ddots \\
0 & \ddots & \ddots & \ddots & \ddots & \ddots \\
\end{bmatrix} - \begin{bmatrix} \tilde{W}^{(1,1)} & \ldots & \tilde{W}^{(1,n)} \\
\vdots & \ddots & \vdots \\
\tilde{W}^{(n,1)} & \ldots & \tilde{W}^{(n,n)}
\end{bmatrix}
\]

\[
\tilde{D}^{(k)} = \sum_{p=1}^{n} D^{(k,p)} \otimes \tilde{D}^{(k,p)}
\]

\[
[D^{(k,p)}]_{i,i} = \sum_{j=1}^{m^{(p)}} [W^{(k,p)}]_{i,j}
\]

\[
\tilde{W}^{(k,\ell)} \equiv W^{(k,\ell)} \otimes W^{(k,\ell)}
\]
ALTERNATIVE EXPLANATION OF THE EQUATION

\[ J(\{f_{ijk}\}) := \frac{\sigma}{2} \sum_{i,j,k,\ell,m,n} w_{ijk,\ell mn} (f_{ijk} - f_{\ell mn})^2 \]
\[ + \frac{1}{2} \sum_{(i,j,k) \in E} (f_{ijk} - f_{ijk}^*)^2 \]

- \( w_{ijk,\ell mn} \) is the symmetric triplet-wise similarity between two triplets \((x_i, y_j, z_k)\) and \((x_\ell, y_m, z_n)\).
- The **first term** indicates that the two link strength values \(f_{ijk}\) and \(f_{\ell mn}\) for the two triplets should be close to each other if the similarity \(w_{ijk,\ell mn}\) between the two triplets is large.
- The **second term** is the loss function that fits the predictions to their target values for the triplets in \(E\) (set of indices for triplets whose link strength is known).
- \(\sigma > 0\) is a regularization parameter which balances the two terms.
**Method - Mathematics**

- We want to minimize that equation -> differentiate

\[ \frac{\partial J}{\partial \text{vec}(F)} = 0 \]

\[ (\sigma L + I)\text{vec}(F) = \text{vec}(A^*) \]

- Solve this with an accelerated version of the conjugate gradient method
  - using vec-trick
    \[ (A_Y \otimes A_X)\text{vec}(B) = \text{vec}(A_X B A_Y) \]
EXPERIMENTS

○ Experiment 1
  ● Non-simultaneous inference with Link Propagation
    ○ Set all cross-species similarity matrices to 0
  ● Simultaneous Link Propagation

○ Experiment 2
  ● Link Propagation
  ● P-SVN, using the same similarity matrices
  ● Baseline: Nadaraya–Watson estimator (kernel regression method)

○ 5 trials
  ● 25%, 50% or 75% of original node pairs for training
EXPERIMENT 1: SIMULTANEOUS VS. INDIVIDUAL INFEERENCE

Table 1. Comparison of individual inferences and simultaneous inference

<table>
<thead>
<tr>
<th>Ratio of training data (%)</th>
<th>Caenorhabditis elegans</th>
<th>Helicobacter pylori</th>
<th>Saccharomyces cerevisiae</th>
<th>Total</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Proposed (individual)</td>
<td>Proposed (individual)</td>
<td>Proposed (individual)</td>
<td>Proposed (individual)</td>
</tr>
<tr>
<td>25</td>
<td>0.702 ± 0.004</td>
<td>0.600 ± 0.007</td>
<td>0.851 ± 0.005</td>
<td>0.749 ± 0.002</td>
</tr>
<tr>
<td>50</td>
<td>0.712 ± 0.005</td>
<td>0.617 ± 0.009</td>
<td>0.901 ± 0.005</td>
<td>0.786 ± 0.005</td>
</tr>
<tr>
<td>75</td>
<td>0.727 ± 0.008</td>
<td>0.629 ± 0.016</td>
<td>0.921 ± 0.008</td>
<td>0.806 ± 0.006</td>
</tr>
</tbody>
</table>

Predictive performances are measured in AUC. Simultaneous inference of multiple species can improve the predictive performance.
**EXPERIMENT 2: LINK PROPAGATION VS. P-SVN**

Table 2. Comparison of the proposed method, the P-SVM and the kernel regression (KR)

<table>
<thead>
<tr>
<th></th>
<th>Caenorhabditis elegans</th>
<th>Helicobacter pylori</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>KR (simultaneous)</td>
<td>P-SVM (simultaneous)</td>
</tr>
<tr>
<td>Ratio of training data (%)</td>
<td></td>
<td></td>
</tr>
<tr>
<td>25</td>
<td>0.593 ± 0.002</td>
<td>0.722 ± 0.007</td>
</tr>
<tr>
<td>50</td>
<td>0.599 ± 0.006</td>
<td>0.752 ± 0.008</td>
</tr>
<tr>
<td>75</td>
<td>0.605 ± 0.012</td>
<td>0.774 ± 0.013</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th></th>
<th>Saccharomyces cerevisiae</th>
<th>Total</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>KR (simultaneous)</td>
<td>P-SVM (simultaneous)</td>
</tr>
<tr>
<td>Ratio of training data (%)</td>
<td></td>
<td></td>
</tr>
<tr>
<td>25</td>
<td>0.822 ± 0.009</td>
<td>0.832 ± 0.007</td>
</tr>
<tr>
<td>50</td>
<td>0.883 ± 0.002</td>
<td>0.884 ± 0.005</td>
</tr>
<tr>
<td>75</td>
<td>0.914 ± 0.006</td>
<td>0.914 ± 0.004</td>
</tr>
</tbody>
</table>

Predictive performances are measured in AUC. The proposed method is competitive with the P-SVM that is one of the state-of-the-art methods.
EXPERIMENT 2: LINK PROPAGATION VS. P-SVN
**Link Propagation**

- simultaneous inference of multiple biological networks
- semi-supervised learning to exploit the information of protein pairs with unknown links
- computational efficiency
- simple and easy to implement

- can be used to infer asymmetric networks
THANK YOU FOR YOUR ATTENTION