Improved Gene Ontology Annotation Predictions through Bayesian Network Post-processing

Marco Tagliasacchi, Marco Masseroli

*Dipartimento di Elettronica e Informazione, Politecnico di Milano, Piazza Leonardo da Vinci 32, 20133 Milano, Italy*
Improved GO Annotation Predictions through Bayesian Network Post-processing

Summary

- Motivation
- Related work
- Problem statement and goal
- SVD method
- Bayesian network method
- Evaluation results
- Conclusions
Motivation

- Several controlled vocabularies and ontologies available and used to functionally annotate genes and proteins
  - Gene Ontology is the most widely used
    - Biological processes
    - Molecular functions
    - Cellular components

- Controlled annotations are paramount to:
  - Support biological interpretation of experimental results
  - Derive new biomedical knowledge
Annotation issues:

- Not exhaustive
  - Only a subset of genes and proteins of sequenced organisms known and annotated
- Incomplete annotations
  - Biological knowledge yet to be discovered
- Incorrect annotations
  - Possibly those inferred from electronic annotations
- Only few reliable annotations
  - By time consuming human curation

Extremely useful computational methods:

- Reliably predict annotations
- Provide prioritized lists of predicted annotations to be checked by curators
Related work

- **Prediction** of annotation profiles has been addressed in the past literature:

  - Methods based on existing annotations:
    - Decision trees/Bayesian networks [Kings et al., 2003]
    - Singular value decomposition (SVD) [Khatri et al., 2005]
    - k-NN classifiers [Tao et al., 2007]
    - ...

  - Methods based on other information sources:
    - Microarray data [Barutcuoglu et al., 2006]
    - Mined textual information [Raychaudhuri et al., 2002], [Perez et al., 2004]
    - ...

Problem statement and goal

- Propose a post-processing method to be applied to the output of the SVD method [Khatri et al., 2005]

- Fix the issue related to the existence of anomalous predictions of ontological annotations:
  - A gene might be predicted annotated to an ontology term, but not to one of its ancestors

GO:0003647 Molecular function
GO:0005215 Transporter activity
GO:0022857 Transmembrane transporter activity
GO:0022804 Active transmembrane transporter activity
GO:0015291 Secondary active transmembrane transporter activity
GO:0022891 Substrate-specific transmembrane transporter activity
GO:0015075 Ion transmembrane transporter activity
GO:0008509 Anion transmembrane transporter activity
Improved GO Annotation Predictions through Bayesian Network Post-processing

Proposed solution

- Leverage the **semantic relationship** between ontological terms as expressed by the ontology structure
- Construct a **Bayesian network** based on the ontology topology and use the output of SVD as prior evidence
- Produce corrected **anomaly free** annotation profiles

GO:0003647 Molecular function
GO:0005215 Transporter activity
GO:0022857 Transmembrane transporter activity
GO:0022804 Active transmembrane transporter activity
GO:0015291 Secondary active transmembrane transporter activity
GO:0022891 Substrate-specific transmembrane transporter activity
GO:0015075 Ion transmembrane transporter activity
GO:0008509 Anion transmembrane transporter activity

Output score of the proposed method

- GO:0003647 (1.00)
- GO:0005215 (1.00)
- GO:0022857 (1.00)
- GO:0022804 (0.07)
- GO:0015291 (0.02)
- GO:0022891 (1.00)
- GO:0015075 (0.99)
- GO:0008509 (0.67)
1. Input: available direct annotations

\[
A = \begin{bmatrix}
0 & 1 & 0 & 0 & 0 & 0 & 1 & 0 & \ldots & 0 \\
0 & 1 & 0 & 0 & 0 & 1 & 1 & 0 & \ldots & 0 \\
0 & 1 & 0 & 1 & 0 & 0 & 0 & 0 & \ldots & 0 \\
1 & 0 & 0 & 0 & 0 & 0 & 0 & 1 & \ldots & 0 \\
\vdots & \vdots & \vdots & \vdots & \vdots & \vdots & \vdots & \vdots & \vdots & \vdots \\
1 & 1 & 0 & 0 & 0 & 0 & 0 & 0 & \ldots & 0
\end{bmatrix}
\]

Ontological terms (e.g. GO terms)

- GO:0003647 Molecular function
- GO:0005215 Transporter activity
- GO:0022857 Transmembrane transporter activity
- GO:0022804 Active transmembrane transporter activity
- GO:0015291 Secondary active transmembrane transporter activity
- \textbf{GO:0022891 Substrate-specific transmembrane transporter activity}
- GO:0015075 Ion transmembrane transporter activity
- GO:0008509 Anion transmembrane transporter activity
2. Annotation unfolding:

\[
\tilde{A} = \begin{bmatrix}
0 & 1 & 1 & 0 & 0 & 0 & 1 & 1 & \ldots & 0 \\
0 & 1 & 1 & 0 & 0 & 1 & 1 & 1 & \ldots & 0 \\
0 & 1 & 1 & 1 & 0 & 0 & 1 & 1 & \ldots & 0 \\
1 & 0 & 0 & 0 & 0 & 0 & 0 & 1 & \ldots & 0 \\
& & & & & & & & & \\
& & & & & & & & & \\
1 & 1 & 1 & 0 & 0 & 0 & 0 & 0 & \ldots & 0
\end{bmatrix}
\]

**Ontological terms (e.g. GO terms)**

- GO:0003647 Molecular function
- GO:0005215 Transporter activity
- GO:0022857 Transmembrane transporter activity
- GO:0022804 Active transmembrane transporter activity
- GO:0015291 Secondary active transmembrane transporter activity
- GO:0022891 Substrate-specific transmembrane transporter activity
- GO:0015075 Ion transmembrane transporter activity
- GO:0008509 Anion transmembrane transporter activity
Improved GO Annotation Predictions through Bayesian Network Post-processing

SVD method

3. Compute SVD:

\[ \tilde{A} = U \Sigma V^T \]

4. Compute reduced rank approximation:

\[ \tilde{A}_k = U_k \Sigma_k V_k^T \]

5. Apply threshold (\(\tau\)):

- If \(\tilde{A}_k(i,j) > \tau\) and \(\tilde{A}(i,j) = 0\) \(\rightarrow\) predicted new annotation (FP)
- If \(\tilde{A}_k(i,j) > \tau\) and \(\tilde{A}(i,j) = 1\) \(\rightarrow\) confirmed annotation (TP)
- If \(\tilde{A}_k(i,j) \leq \tau\) and \(\tilde{A}(i,j) = 0\) \(\rightarrow\) confirmed no annotation (TN)
- If \(\tilde{A}_k(i,j) \leq \tau\) and \(\tilde{A}(i,j) = 1\) \(\rightarrow\) annotation to be checked (FN)
The output of the SVD method might contain anomalous predictions

The real valued output of the SVD method might be such that:

$$\tilde{A}_k(i,j) > \tilde{A}_k(i,r)$$

where $r$ is ancestor of $j$

After thresholding, term $j$ might result annotated to gene $i$, while term $r$ is not
Improved GO Annotation Predictions through Bayesian Network Post-processing

Bayesian network method

- Design a Bayesian network to remove anomalous predictions
  - Input: real-valued scores computed by SVD method
  - Output: anomaly-free real-valued scores

- Bayesian network structure based on **ontology topology**
  - Term nodes
  - Evidence nodes

- Need to define **conditional probabilities**
Improved GO Annotation Predictions through Bayesian Network Post-processing

**Bayesian network method**

For each gene i:

- **Term nodes (t-nodes) conditional probabilities**

  \[
  p_i(t_j | t_{c_1}, t_{c_2}, \ldots, t_{c_L})
  \]

  Estimated from available annotations

\[
\begin{array}{c|c|c}
  t_j & t_{c_1}, t_{c_2}, \ldots, t_{c_L} \\
  \hline
  1 & 1 0 \ldots 0 \\
  1 & 0 1 \ldots 0 \\
  1 & 1 1 \ldots 0 \\
  \vdots & \vdots \ldots \vdots \\
  1 & 0 0 \ldots 0 \\
  0 & 1 0 \ldots 0 \\
  0 & 0 1 \ldots 0 \\
  \vdots & \vdots \ldots \vdots \\
  0 & 0 0 \ldots 0
\end{array}
\]
Evidence nodes (e-nodes) conditional probabilities:
- Gaussian Mixture Model (estimated from available \(<t_j,e_j>\) pairs)
Improved GO Annotation Predictions through Bayesian Network Post-processing

Bayesian network method

- For each gene $i$, e-nodes are fed with the real-valued output of the SVD method

- Inference (junction tree algorithm) is performed to get the a-posteriori marginal distribution $p_i(t_j) \ \forall i, j$ of the binary-valued t-nodes:
  - The probability of gene $i$ to be annotated to term $j$
The a-posteriori marginal distribution $p_i(t_j)$:

- Provides a real-valued output to be used for producing a ranked list of candidate annotations
- Can be thresholded, similarly to the output of the SVD method, but without anomalies

Output score of the proposed Bayesian network method

Fixed anomaly

GO:0003647 Molecular function
GO:0005215 Transporter activity
GO:0022857 Transmembrane transporter activity
GO:0022804 Active transmembrane transporter activity
GO:0015291 Secondary active transmembrane transporter activity
GO:0022891 Substrate-specific transmembrane transporter activity
GO:0015075 Ion transmembrane transporter activity
GO:0008509 Anion transmembrane transporter activity
Improved GO Annotation Predictions through Bayesian Network Post-processing

Evaluation results

- Tested on:
  - *Saccharomyces cerevisiae* (SGD) and *Drosophila melanogaster* (FlyBase)
  - Gene Ontology annotations (Oct 2008)
    - Biological Processes (BP)
    - Molecular Functions (MF)
    - Cellular Components (CC)
  - Retaining only terms used to annotate at least 10 genes

<table>
<thead>
<tr>
<th></th>
<th>BP</th>
<th>MF</th>
<th>CC</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Genes</td>
<td>Terms</td>
<td>Genes</td>
</tr>
<tr>
<td>SGD</td>
<td>5,351</td>
<td>807</td>
<td>4,329</td>
</tr>
<tr>
<td>FlyBase</td>
<td>6,731</td>
<td>1,084</td>
<td>6,907</td>
</tr>
</tbody>
</table>

- Results presented for GO Molecular Functions of SGD
- Similar conclusions for FlyBase and other GO ontologies
Observations:

- The total number of $FP + FN$ is similar in the two methods (SVD and BN).
- The SVD method produces a large number of anomalies when the threshold ($\tau$) is close to 0 or 1.
- The Bayesian network (BN) post-processing removes all anomalous annotation predictions.
FP and anomaly rates
- By dividing both anomaly and FP counts by number of total original negative annotations (i.e., FP+TN)
  - FP rate = 0.01 \( \rightarrow \) 11% of predicted annotations
  - FP rate = 0.005 \( \rightarrow \) 7.5% of predicted annotations
  - FP rate = 0.001 \( \rightarrow \) 1.8% of predicted annotations

- SVD method: anomalous annotation predictions:
  - FP rate = 0.01 \( \rightarrow \) 11% of predicted annotations
  - FP rate = 0.005 \( \rightarrow \) 7.5% of predicted annotations
  - FP rate = 0.001 \( \rightarrow \) 1.8% of predicted annotations

- Bayesian network method: anomalies are always zero
Proposed a post-processing method to remove anomalous annotation predictions produced by SVD method

The proposed method:

- Provides a ranked list of probable annotations consistent with the ontology structure
- Not only avoids anomalous annotation predictions, but also improves predictions globally, thus busting performance of computational method using them
- Is not bounded to GO, but it is applicable to any ontological annotations

Possible further annotation predictions improvement:

- By separately estimating term co-occurrences for each functionally consistent cluster of genes