Protein Function Prediction by Integrating Different Data Sources

Liang Lan, Nemanja Djuric, Yuhong Guo, Slobodan Vucetic

Temple University, Philadelphia, USA

AFP/CAFA, Vienna, Austria, July 15-16th, 2011

Introduction

- Protein function annotation - key challenge in post-genomic era
- Experimental annotation accurate, but slow and expensive
- Large amount of information available
- Data mining techniques can help when dealing with available, large-scale data sets

Our approach

- Integrate information from different sources to predict gene functions
- We consider following data sources:
  - Protein sequence similarity
  - Protein-protein interaction
  - Gene expression
- Hypothesis: Including information from various sources results in better predictor performance

Methodology

- We use weighted k-Nearest Neighbour algorithm to calculate likelihood that protein p has function f

$$score(p, f) = \sum_{p' \in N_k(p)} sim(p, p') \cdot I(f \in functions(p'))$$

- Simple to implement, yet competitive when compared to SVM
- Different $sim(p, p')$ can be obtained with different data sources

Methodology - Cont'd

- We calculated different scores using different sources (sequence similarity, PPI, gene expression) for each (p, f) pair
- Total of J gene expressions, resulted in J+2 scores that are combined:

$$\begin{aligned} score(p,f) = {} & w^{SEQ} \cdot score^{SEQ}(p,f) + \\ & w^{PPI} \cdot score^{PPI}(p,f) + \\ & \sum_{j=1}^{J} (w_j^{EXP} \cdot score_j^{EXP}(p,f)) \end{aligned}$$

Integrating different scores

- How to find weights w^{SEQ}, w^{PPI}, w_j^{EXP}?
- We considered several methods:
  - Assigning the same weights to all scores
  - Weight optimization by likelihood maximization
  - Weight optimization by large margin approaches
- Also considered enhancing similarity scheme using approach from Pandev *et al*.*

* “Incorporating functional inter-relationships into protein function prediction algorithms”, BMC Bioinformatics (2009)

Max-margin approach

- Define the following optimization problem:
  - Given n genes and m scores, and $f(x, y)$ is an m x 1 vector of scores for gene x and function y, solve:

$$\min_{\mathbf{w},\xi} \frac{1}{2} \|\mathbf{w}\|^2 + C \sum_i \sum_{y \in Y_i, \bar{y} \in \bar{Y}_i} \xi_i(y, \bar{y})$$

$$\text{s.t. } \mathbf{w}^{\mathrm{T}}(f(x_i, y) - f(x_i, \bar{y})) \geq 1 - \xi_i(y, \bar{y}), \ \forall i, y \in Y_i, \bar{y} \in \bar{Y}_i$$

$$\xi_i(y, \bar{y}) \geq 0, \ \forall i, y \in Y_i, \bar{y} \in \bar{Y}_i$$

where $\mathbf{w}$ is an m x 1 weight vector learned during training, and C is a regularization parameter

Experimental setup

- We focused on function prediction for human proteins
- Data sources:
  - Sequence similarity scores for all pairs of CAFA proteins
  - Gene expressions - 392 Affymetrix GPL96 Platform microarray data sets from GEO
  - PPI - Physical interactions between human proteins listed in OPHID database

Experimental setup - Cont’d

- 8,714 annotated human proteins in CAFA training set
- Out of those 8,714, total of 2,869 proteins covered by all three data sources
- For evaluation, only GO functions annotated by more than 10 proteins are considered
  - This resulted in 240 MF and 1,123 BP GO terms
- Neighbourhood size fixed to 20

Score averaging scheme

- None of the considered approaches worked significantly and consistently better than simple averaging
- As a result, we give the same weight to all 3 data sources:

$$w^{SEQ} = w^{PPI} = 1/3$$

$$w_j^{EXP} = 1/(3J)$$

Results (average AUC)

- *ver. 1* - neighbors found among only 2,869 overlapping human proteins
- *ver. 2* - neighbors found among all 8,714 human proteins
- *ver. 3* - neighbors found among all 36,924 CAFA training proteins

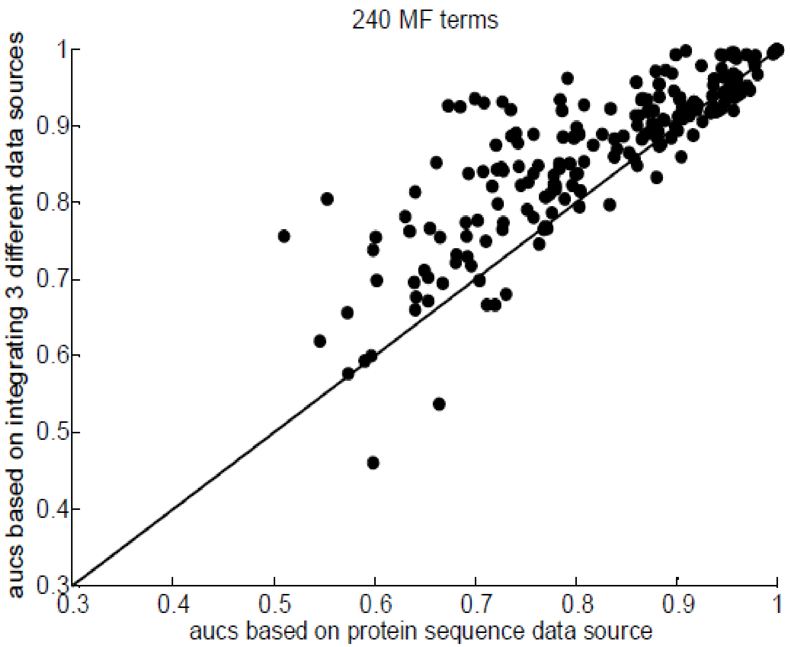
Data source	**MF terms**	**BP terms**
Microarray data	0.6442	0.6279
PPI data	0.6283	0.6671
Protein Sequence data, ver. 1	0.7636	0.6642
Protein Sequence data, ver. 2	0.7896	0.6921
Protein Sequence data, ver. 3	0.8396	0.7537
Integrating 3 data sources, ver. 1	**0.8134**	**0.7468**
Integrating 3 data sources, ver. 2	**0.8494**	**0.7939**
Integrating 3 data sources, ver. 3	**0.8788**	**0.8165**

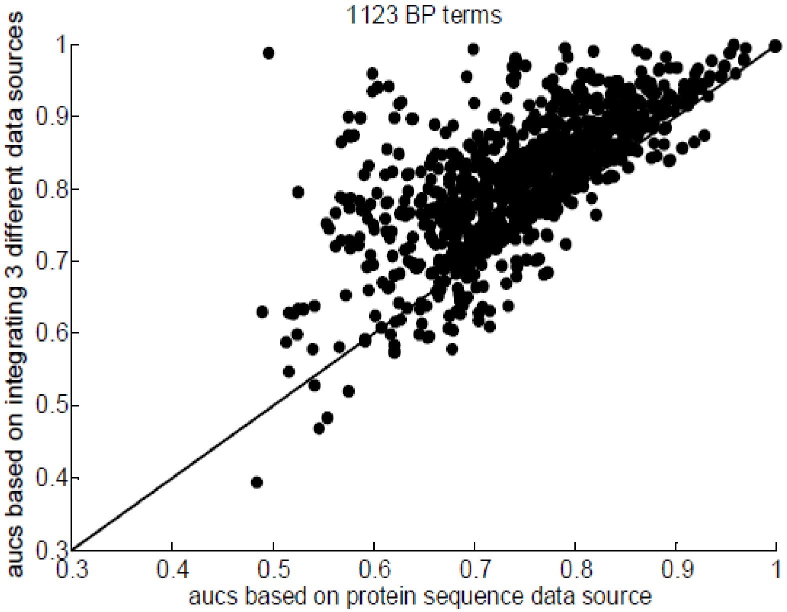
Discussion

- Several important conclusions arise:
  - Gene expression is more useful for MF, while PPI is more useful for BP prediction
  - Sequence similarity data is superior to both gene expression and PPI data
  - It is beneficial to transfer functions to human proteins from their orthologues
  - Integration of data sources improves AUC significantly for both MF and BP terms

Results - Cont'd

- Comparison of AUC of sequence similarity scores (*ver*. 3) and integrated scores (*ver*. 3) for each GO term





Conclusion

- Some sources are more beneficial for BP, while some for MF terms prediction
- Integration of different sources improves function prediction significantly
- Exploring new integration techniques could lead to even better results

Thank you!

Questions?