Poster J-29

Protein Interaction Prediction with Supervised Learning Methods

Authors:
Oksana Kohutyuk (*Department of Comp. Science, Iowa State University*)
Vasant Honavar (*Department of Comp. Science, Iowa State University*)

Short Abstract: We evaluated the effects of class distribution in training and test data in protein interaction prediction on performance of different classifiers (DT, NB, SVM, RF), and developed a recursive algorithm that utilizes predictions made at a given iteration for future classifications, which showed improved performance on small data sets.

Long Abstract:

Protein Interaction Prediction
with Supervised Learning Methods

Protein-protein interactions are a vital part of every biological process, and elucidating protein interactions is crucial to advancing our understanding of the biochemical processes within the cell. Proteins serve a vital role in many major cellular processes, and anomalous protein interactions disrupt the normal cascade of biological processes and may even lead to diseases such as Alzheimer's. Discovering protein interactions can also assist us in making predictions about functions of unknown proteins, based on the premise that the function of unknown protein may be dependent on the function of the proteins it interacts with. Several high-throughput experimental approaches, such as yeast two-hybrid (Y2H) and mass spectrometry methods have been developed to determine protein interactions, but results of these methods are often incomplete and inaccurate, exhibiting high false positive and false negative rates, and, when combined, yield a small overlap in predicted interactions (von Mering et al). Recently, various supervised learning techniques (Qi et al, Lu et al, Jansen et al and others) were applied to integrate proteomic and genomic data and utilize existing knowledge of protein-protein interactions to predict unknown protein-protein interactions. Such methods complement current experimental approaches and the predictions obtained can serve as a basis for experimental refinement of interactome discovery.

This study evaluated the performance of four different classifiers (Decision Trees, Naïve Bayes, Random Forest, and Support Vector Machines), trained on small data samples and combined through bagging, on protein interaction prediction using four genomic features on a dataset with extremely skewed distribution. A recursive classifier that exploits predicted relationships for future classification was developed and shown to have an increased prediction accuracy compared to traditional classifiers on small data sets.

A subset of data from Lu et al. (see http://networks.gersteinlab.org/BayesFeatures/) was used for this study. A positive gold standard dataset of protein co-complex membership, extracted from MIPS, and the negative dataset, constructed by pairing proteins from different subcellular compartments, was used to assess predictive power of various classification schemes. Four features with the top predictive power (according to Lu et al) were selected

for training: mRNA co-expression, protein co-essentiality, MIPS functional similarity, and GO functional similarity. In contrast to the study by Qi et al, which used 162 features, many of which are not readily available for organisms other than yeast, we decided to use a minimal number of basic proteomic features to investigate the effects of data distribution and classification schemes on prediction accuracy and attempt to improve prediction results using these features only.

We ran a 10-fold cross-validation on the original dataset with varying parameters. To address the issue of data size and data distribution, which impacts the running time and performance of most classifiers, n base classifiers (either Decision Trees, Naïve Bayes, SVM, or Random Forest) were trained on small randomly constructed subsets of training data, and for each new instance x to be classified, every base classifier outputted a class prediction and the results were combined though majority voting. We measured the effects of both training data distribution and test data distribution on prediction accuracy for each classifier type.

Our results indicated that whereas different classification schemes and data distributions yielded similar ROC curves with .92-.97 area, the prediction-recall curves differed tremendously for various classification schemes and data distributions. In general, Random Forest produced the best predictions, which is consistent with the results from Qi et al study. RF proved to be largely insensitive to data distribution and produced better precision-recall curves than other methods on the extremely skewed distribution, even in cases where RF did not have the highest ROC area. Deliberately sampling test data to have an equal positive-negative class distribution had a very strong positive impact on prediction results, thus illustrating the importance of class distribution in unlabeled data in the prediction task. Class distribution in training samples was also varied, changing from the original very skewed towards negative class distribution to a completely balanced distribution. Surprisingly, training classifiers on the samples with original distribution produced the best precision-recall curves when compared to classifiers trained on more balanced samples.

To improve classification performance, an iterative algorithm, RPIP (Recursive Protein Interaction Prediction) was designed, where at each iteration the averages of corresponding similarity measures between the first protein in a target protein pair and the proteins predicted to interact with the second protein in the pair in the previous iteration (its neighbors), and the average similarities between the second protein and the first protein's neighbors, are used to update the classifier during training. To make a prediction for a new instance, a similar procedure is followed. The reasoning behind a recursive classifier that utilizes predictions made at a given iteration to make future predictions is that protein-protein interactions are not totally isolated and independent of each other, and interaction partners of one protein in the interacting pair often have an effect on the other protein's connections. Proteins often work together in multi-protein complexes, and it is very likely that two proteins in the same complex might have common interaction partners.

Preliminary results showed improvement in prediction results of RPIP comparing to the other four methods on very small datasets, and RPIP seems to be a promising new approach in protein interaction prediction. Some optimization steps need to be taken to improve the running time of the algorithm, and it still remains to be tested on large datasets.

References:

Lu et al. Assessing the limits of genomic data integration for predicting protein networks. Genome Research 2005: 15:945-953.

von Mering et al. Comparative assessment of large-scale data sets of protein–protein interactions. Nature 2002;417:399–403.

Qi et al. Evaluation of different biological data and computational classification methods for use in protein interaction prediction. PROTEINS. 2006: 63(3):490-500.