Critique

Alexey Dievsky

November 24, 2012

Abstract

Probabilistic prediction and ranking of human protein-protein interactions, Michelle S. Scott and Geoffrey J. Barton, 2007

The paper presents a new method for predicting novel protein-protein interactions based on the existing studies. The method employs a naïve Bayes classifier which combines several predictors trained from the sample distributions.

Summary

The construction of the interactome (the map of protein-protein interactions) is an important area of research in modern biology, as the interactions between the proteins define virtually all activity of a cell and, consequently, of the whole organism. Multiple disorders are caused by an interaction malfunction, thus correctly identifying the interaction in question is often the first step of a cure research. The knowledge of the interactions leads to better understanding of the biological pathways.

This said, currently there is no reliable method of imputing the protein-protein interactions. All the attempts at *de novo* discoveries are based on the indirect evidence, such as a frequent co-occurrence of the two proteins. Due to the unreliable and stochastic nature of such prediction methods, either the FDR (false discovery rate) becomes unreasonably high (many incorrect interactions are imputed), or the power of the method becomes unreasonably low (no correct interactions are discovered). While the potential interactions may be verified manually through different means (e.g. the yeast two-hybrid system), the process is costly and time-consuming.

The authors propose a new method for this task which essentially combines several well-known predictors using a naïve Bayes approach:

$$\begin{aligned}\frac{P(I|f_1,\ldots,f_n)}{P(\neg I|f_1,\ldots,f_n)} &= \frac{P(f_1,\ldots,f_n|I)\cdot P(I)}{P(f_1,\ldots,f_n|\neg I)\cdot P(\neg I)} \\ &= \frac{P(f_1|I)}{P(f_1|\neg I)}\cdot\ldots\cdot\frac{P(f_n|I)}{P(f_n|\neg I)}\cdot\frac{P(I)}{P(\neg I)}\end{aligned}$$

Each predictor produces an independent LR (likelihood ratio) value, and all the LRs are then multiplied together with a prior LR. The result can be compared against a fixed threshold value to decide if the interaction is predicted between the given pair of the proteins.

Critique

Despite the competitive performance of the presented method, the scientific basis of the underlying model remains questionable.

First of all, the use of naïve Bayes approach is not given as much attention as it requires. This approach corresponds to the reality if and only if the individual predictors are altogether independent. Virtually no elaboration is provided on that issue, except the mention of calculating the Pearson correlation coefficient for each pair of the predictors and deeming it low enough to proclaim approximate independence. However, the Pearson correlation coefficient is not a defining criterion of independence; while the true correlation of the independent values is always zero, the opposite statement is not valid. Even if we agree to such use of the correlation coefficient, it will only infer the pairwise independence, while for the naïve Bayes approach to work, the predictors must be mutually independent. No evidence supportive of the mutual independence is provided.

The selection of the prior LR is given some consideration, only to conclude that we have no sound means of validating our choice of $\frac{1}{400}$, except that it seems to work. Other prior LRs considered vary considerably.

Further, for a method aiming at a scientific validity, the sentences like "All preliminary scores above 10 were kept. This parameter was determined empirically." seem strange.

The choice made by the authors: "When a particular state of a feature occurs only in positive examples (known interacting proteins), the likelihoods are set to the highest non-infinite value of any state for that feature (to avoid infinite values)." has no obvious advantages (and several obvious disadvantages) against using any sort of smoothing, which is not considered at all.

The negative training dataset was mostly produced by taking a random set of protein pair and removing the pairs known to interact. This procedure produces not a dataset of non-interacting proteins, but rather a dataset of proteins not known to interact. While this is probably the best we can do for the time, the effects of this bias are not taken into account in any way.

We conclude by noting that despite the mentioned flaws, the method most likely actually works (as verified by different means). What seems to be lacking is the thorough elaboration on the correspondence between the model and the reality.