Label-Space Dimensionality Reduction and a Similarity-Based Representation for Protein Function Prediction

Makrodimitris, S.[1,2,]*, van Ham, R.C.H.J.[1,2], Reinders, M.J.T.[1]
[1]Delft Bioinformatics Lab, Delft University of Technology, Mekelweg 4, 2628CD, the Netherlands
[2]Keygene N.V., Agro Business Park 90, 6708PW, the Netherlands
*To whom correspondence should be addressed: s.makrodimitris@tudelft.nl

1. INTRODUCTION

Machine Learning (ML) is used in many bioinformatics problems, but it is not so widely applied in Automatic Function Prediction (AFP). One of the reasons for this is that ML techniques require a feature representation of proteins which is not trivial to obtain. We propose a novel feature representation of a protein based on its sequence similarity to a *set* of annotated training proteins. Additionally, the CAFA challenges (1) have pointed out that BLAST performs on par with, or even better than, most of the more elaborate participating methods in the Biological Process (BP) and Molecular Function subcategories of the Gene Ontology (GO). We argue that this apparent "failure" of many AFP algorithms is partly due to the nature of the GO itself. Although very intuitive, GO terms seem to be hard for computers to learn. A reason for this might be that GO contains many terms, in particular in the BP ontology (>29,000), which is also the "hardest" category (1). We propose to use Label-Space Dimensionality Reduction (LSDR) techniques to transform GO terms into a more compact latent representation that is easier to predict.

2. MATERIALS & METHODS

Given N training proteins, we use X to denote an NxN matrix where X_{ij} contains the percent sequence identity between the i-th and the j-th protein. We also define an NxL label matrix Y, where L is the number of GO terms in an ontology and $Y_{ij} = 1$ if the i-th protein is annotated with the j-th term and zero otherwise.

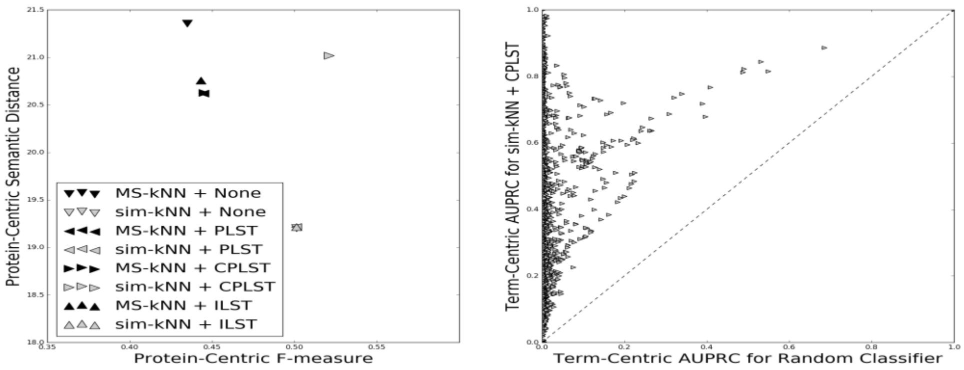
Using LSDR, we project Y into a lower-dimensional representation Y'. We then train a k-Nearest Neighbor (kNN) regressor from X to Y'. For each test protein, the similarities to *all* training proteins are calculated and fed into the regressor to predict Y', i.e. each protein is represented by its similarities to the training proteins (hence the similarity representation). We then apply the inverse LSDR transform to obtain a score for every GO term being associated with that protein.

We test three linear LSDR methods, namely *Principal Label Space Transformation* (PLST) (2), *Conditional Principal Label Space Transformation* (CPLST) (3) and a new one, *Independent Label Space Transformation* (ILST). Briefly, when the desired number of dimensions is m, the PLST transformation matrix consists of the singular vectors of Y corresponding to the m largest singular values. The CPLST matrix consists of the singular vectors of $Y^T XX^+ Y$ corresponding to the m largest singular values, where X^+ denotes the left pseudo-inverse of X. ILST obtains m independent target labels by applying *Independent Component Analysis* (ICA) to Y.

We compared our similarity-based predictor (*sim-kNN*) to the sequence-based *MS-kNN* algorithm (5), in which the k proteins with the highest sequence similarity to a test protein are used in a weighted voting scheme to calculate scores for a every protein-term pair. The parameters of each method were tuned in a 3-fold double-loop cross-validation scheme (4) in which the inner loop identifies the optimal parameters and the outer loop evaluates performance on unseen data.

3. RESULTS

Figure 1a shows the cross-validation performance of the combinations of the two algorithms (*sim-kNN* and *MS-kNN*) with the 3 LSDR methods, as well as without any LSDR (*None*). Both metrics were calculated per test protein and then averaged across proteins (protein-centric evaluation). Figure 1b shows the Area Under the Precision-Recall Curve (AUPRC) for the *sim-kNN* method coupled with CPLST compared to a random classifier, to get an impression of the improvement over random classification per term. For the vast majority of GO terms, the performance is considerably higher than what expected by chance.



(a) (b)

Figure 1: (a) Protein-Centric F-measure (x-axis) vs Semantic Distance (y-axis) of all methods measured with 3-fold cross-validation. In gray, our method (sim-kNN) and in black, the sequence-based MS-kNN. Marker shapes denote the LSDR methods. (b) Term-Centric AUPRC of the sim-kNN method with CPLST (y-axis) vs the AUPRC expected by a random classifier (x-axis). Dashed line shows the y = x line.

4. DISCUSSION

We propose a feature representation of proteins that has not been used for AFP before, in which a protein is described by its similarity to a set of training proteins. This generates a feature matrix, enabling the use of ML approaches. As a proof of concept, we use sequence identity as a similarity measure, but it is straightforward to extend the method to other data types, such as gene co-expression and protein-protein interactions. When coupled with *kNN* regression, this representation outperforms sequence-based *MS-kNN* (5) in both protein-centric metrics used.

Furthermore, we propose to transform GO terms to reduce the number of terms by exploiting the redundancies between them. We tested two previously published LSDR methods, PLST (2) and CPLST (3), as well as a novel approach that uses ICA to create uncorrelated target labels. Our results show that LSDR methods either slightly improve or do not affect the performance of both AFP methods tested, while reducing the number of target variables by 60-90%.

5. REFERENCES

1. Jiang, Y., Ronnen Oron, T., et al. 2016. An expanded evaluation of protein function prediction methods shows an improvement in accuracy. *Genome Biology* 17:184.

2. Tai, F. and Lin, H.T. 2014. Multilabel Classification with Principal Label Space Transformation : *IEEE Neural Computation* 24:9, pp. 47-58.

3. Chen, Y.N. and Lin, H.T. 2012. Feature-aware Label Space Dimension Reduction for Multi-label Classification. *Advances in Neural Information Processing Systems 25*.

4. Simon R. and Varma S. 2006. Bias in error estimation when using cross-validation for model selection. *BMC Bioinformatics* 7:91.

5. Lan L., Djuric N., Guo Y. and Vucetic S. 2013. MS-kNN: protein function prediction by integrating multiple data sources. *BMC Bioinformatics* 14 Suppl 3:S8.