

# Predicting Gene Ontology Annotations from Sequence Data using Kernel-Based Machine Learning Algorithms

J. J. Ward, J. S. Sodhi, B. F. Buxton, D. T. Jones  
Department of Computer Science  
University College London  
Gower Street, London, WC1E 6BT, UK  
j.ward@cs.ucl.ac.uk

## Abstract

*In this early part of the post-genomic era, the inference of the functions associated with gene products is a necessary first step in understanding the development and maintenance of living cells. We describe the development of a machine learning method for predicting biological process as defined by the Gene Ontology (GO). The algorithm uses features that can be generated from amino acid sequence alone, and does not require further experimental studies such as microarrays, 2-hybrid screens or systematic 'pull-down' assays. The budding yeast *Saccharomyces cerevisiae* is used because of its comprehensive set of functional annotations, but the approach is sufficiently general for application to other eukaryote genomes. The input data include phylogenetic profiles, which represent the distribution of orthologous proteins in the genomes of other organisms, position-specific scoring matrices, and secondary structure and dynamic disorder predictions. These are encoded using diffusion kernels, which are used to represent pair-wise relationships such as sequence or secondary structure element similarity between nodes (proteins) in a graph. These kernels are benchmarked on the process prediction problem using a maximal margin (SVM) learning algorithm.*

## 1. Introduction

The explosive growth in the quantity of sequence data that has occurred in the previous decade has provided the stimulus for automatic techniques that can be used to annotate the structure and function of biological macromolecules. While structural genomics projects have succeeded in mapping much of fold space, our knowledge of the complex processes that co-ordinate the development and maintenance of living cells is far from complete. The obvious first step in improving this under-

standing is to annotate simple functional properties of proteins such as their subcellular locale, their interactions with ligands and other proteins, and their involvement in a particular biochemical pathway.

The various available data sources are each useful for inferring different aspects of protein function. This work focusses on predicting the biological process category of the Gene Ontology [2] for proteins in *Saccharomyces cerevisiae*, and relies on several related data sources that can be recovered directly from amino acid sequence. These sources make use of modular properties of metabolic pathways at the genome level and functional domains at the level of single proteins.

## 2. Data

The data comprises the January 2004 set of yeast sequences and annotations from the *Saccharomyces* Genome Database [1].

Phylogenetic profiles were generated by running a PSI-BLAST search for each yeast protein against a database of 92 complete genomes which was filtered to remove low complexity and transmembrane regions. Each feature in the profile is proportional to the log expectation value of the nearest homologue recovered from each genome.

Secondary structure predictions were obtained using PSIPRED [3] and used to generate a matrix of pair-wise structural similarity relationships using a global alignment of predicted secondary structure elements [5]. A matrix of sequence similarity relationships was obtained using gapped BLAST.

## 3. Algorithms

The terms in the biological process ontology that described more than 50 yeast proteins were used as classes in each comparison. A standard 2-norm soft margin SVM classifier was used to discriminate between proteins in each

process class and all other proteins apart from those annotated with ‘process unknown’. Asymmetric costs for breaching the margin were placed on the each class to account for the unbalanced data sets.

Linear kernels were used to classify proteins using phylogenetic profiles and a combination of amino acid and the predicted structural composition of each protein. Diffusion kernels were generated on the graphs of sequence and structural similarity relationships [4]. These kernels are calculated by taking the matrix exponential of the Laplacian matrix  $L$  on a graph

$$K = e^{\beta L} \quad (1)$$

where  $\beta$  is a parameter analogous with conductivity in physical systems. Differentiation of equation 1 with respect to  $\beta$  provides an equation with a similar form to diffusion or heat equations.

$$\frac{d}{d\beta} K = LK \quad (2)$$

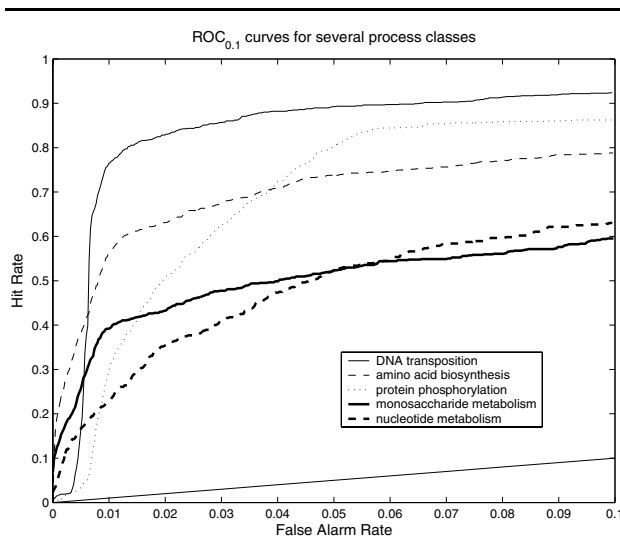
Only the ten highest-scoring structural homologues of each protein were included in the graph of structural relationships as the absolute scores have been shown to correlate poorly with the correct fold, although the rankings for each query protein have been shown to be fairly accurate [5].

## 4. Results

The results were obtained by performing a 3-fold cross-validation on the 4125 yeast proteins with 10 random partitions of the data set. The outputs of each replicate were used to generate ROC curves up to a false positive rate of 0.1 with the area under the curve used as the measure of performance (see figure 1).

Short summaries of the results from each data source are listed below

- The highest accuracies obtained from phylogenetic profile data are attained on biosynthetic and metabolic processes, which suggest that proteins in these pathways tend to be inherited as intact modules.
- Amino acid and structural composition are also predictive of metabolic processes such as membrane lipid synthesis, DNA recombination and ribosome biogenesis although these tend to be associated with specific cellular compartments.
- Results from the diffusion kernel on the set of homology relationships show only slightly higher accuracy than simply transposing the same function as the closest homologue in the BLAST search. However, further optimization of the “conductivity” parameter may yield an improved performance.



**Figure 1. Averaged ROC curves for linear SVMs trained on phylogenetic profiles. The bottom curve represents a random class assignment.**

- At present the kernel on the structural similarities is lower than the other sources, although this may be improved by developing a new scoring scheme for the structural alignments.

## 5. Further Work

In addition to improving results from diffusion kernels, it is intended that the results from the binary classifiers will be integrated into an overall prediction for biological process.

## References

- [1] S. S. Dwight *et al.* *Saccharomyces* Genome Database (SGD) provides secondary annotation using the Gene Ontology (GO). *Nucl. Acids. Res.*, 30(1):69–72, 2002.
- [2] Gene Ontology Consortium. Creating the gene ontology resource: design and implementation. *Genome Research*, 11:1425–1433, 2001.
- [3] D. T. Jones. Protein secondary structure prediction based on position-specific scoring matrices. *J. Mol. Biol.*, 292:196–202, 1999.
- [4] R. I. Kondor and J. Lafferty. Diffusion kernels on graphs and other discrete structures. In *Proceedings of the ICML*, 2002.
- [5] L. J. McGuffin and D. T. Jones. Benchmarking protein secondary structure prediction for protein fold recognition. *Proteins*, 52:166–175, 2003.