Pacific Symposium on Biocomputing 7:374-385 (2002)

PREDICTING THE SUB-CELLULAR LOCATION OF PROTEINS FROM TEXT USING SUPPORT VECTOR MACHINES

B.J. STAPLEY [a], L.A. KELLEY, M.J.E. STERNBERG [b]
Biomolecular Modelling Laboratory, Imperial Cancer Research Fund, 44 Lincoln's Inn Field, London. WC2A 3PX, United Kingdom.
(b.stapley, l.kelley, m.sternberg)@icrf.icnet.uk

We present an automatic method to classify the sub-cellular location of proteins based on the text of relevant medline abstracts. For each protein, a vector of terms is generated from medline abstracts in which the protein/gene's name or synonym occurs. A Support Vector Machine (SVM) is used to automatically partition the term space and to thus discriminate the textual features that define sub-cellular location. The method is benchmarked on a set of proteins of known sub-cellular location from *S.cerevisiae*. No prior knowledge of the problem domain nor any natural language processing is used at any stage. The method out-performs support vector machines trained on amino acid composition and has comparable performance to rule-based text classifiers. Combining text with protein amino-acid composition improves recall for some sub-cellular locations. We discuss the generality of the method and its potential application to a variety of biological classification problems.

1 Introduction

The sub-cellular localisation of a protein is a key element in understanding its function. In order to carry out its physiological role, a protein must be often be proximal to other components involved in that process; thus knowledge of sub-cellular localization can restrict the number of possible pocesses with which a protein can be involved. Location can also alter the experimental approach to characterising a protein - e.g. purification.

Despite the importance of a protein's sub-cellular localisation, automatic prediction or extraction of this property has proved a suprisingly difficult task[1]. It has been know for sometime that the amino acid composition of protein can be an indicator of its sub-cellular location [2]. It is also clear that many cellular compartments have proteins assigned to them according to targeting signals within the protein sequences; however, such signals are not universal or necessarily clearly defined.

[a] present address; Biomolecular Sciences, University of Manchester Institute of Science Technology, PO Box 88, Manchester, UK, M60 1QD
[b] present address; Department of Biological Sciences, Imperial College of Science, Technology and Medicine, London, SW7 2AY, United Kingdom

An alternative approach, pioneered by Eisenhaber and Bork is to use the existing textual information relevant to a protein to classify it to a particular sub-cellular location [3]. This is achieved using a set of manually generated biological rules and the SWISS-PROT annotations of the proteins [4]. After tokenizing the annotations the rules are applied and a sub-cellular location extracted. They named this method Meta-Annotator. The authors report that 88% of SWISS-PROT entries can been assigned to a cellular compartment by this method. This compares very favourably with the 22% that is achieved by simple matching of relevant keywords within the documents.

Despite the success of this technique, it has two inherent weaknesses: first, a set of rules must be generated - this is obviously less intensive than manually classifying the documents, but is subjective and costly in time; second, in order to tokenize the documents they must already be structured - free text cannot be treated in such a manner without recourse to natural language parsing (NLP). NLP is beginning to show great promise within the field of biological informatics and has been successfully applied to extracting protein-protein interactions [5], metabolic pathways [6], and drug/gene relationships [7] from biological text. Although NLP often achieves very good precision, recall is often disappointing - problems of synonymy and polysemy are very difficult to overcome. In this work we investigate whether a simpler approach to the problem can be successfully applied.

The method described in this paper is to treat the protein as a vector of terms from relevant Medline documents. This approach derives from the vector-based model common in information retrieval [8]. The term weights of a vector are a functions of their frequencies within the document collection as a whole and the frequency within the relevant documents. Given a set of protein term-vectors the task is to find some function that partitions the space according to the localisation of the protein. For this task we employ support vector machines (SVM) [9].

Support vector machines are a mathematical method for performing simultaneous dimension reduction and binary classification [9]. SVMs have been applied to the problems of pattern recognition [10], regression estimation [10] and information retrieval [11,?]. Because SVMs cope well with high dimensionality and are very fast to train, they are particularly suited to problems in text data-mining/information retrieval. Kwok studied the use of SVMs in text catagorization of Reuters newswire documents [12]. In this paper, we apply an analogous approach to Medline/SWISS-PROT documents.

We evaluate the performance of SVMs in classifying a set of proteins of known sub-cellular locations from *S. cerevisiae*. Text relevant to these proteins is obtained from Medline by key-word matching of the gene naming terms.

SVMs trained on the resulting term vectors classify the proteins with good precision and recall. We also show that SVMs trained on amino acid compositions are out-performed by our SVMs trained to text data and that combining amino acid composition and term vectors can enhance classification for some sub-cellular locations.

2 Methods

2.1 Document and term processing

To obtain term vector representations of *cerevisiae* proteins we employed the following procedure. First, we scanned 22517 Medline documents for occurrences of yeast gene naming terms. These terms and synonyms were obtained from the Saccharomyces Genome Database gene registry [13] [c]. For each protein, any document that contained an occurrence of the gene name or aliases of that gene was considered relevant. This resulted in a collection of 12596 documents. We employed stop word removal, stemming and removed stemmed terms that occurred in fewer than five documents. The term representation of a gene is a function of the number of relevant Medline documents and the occurrence statistics of the terms. We employed a variant of inverse document frequency (IDF) that takes account of the number of Medline documents relevant to a particular gene. The weight of term i for gene k is given by :

$$\log\left(1+\sum_{j} f_j(w_i)\right) - \log N(w_i) - \log\left(1+R_k\right) \qquad (1)$$

where $f_j(w_i)$ is the frequency of term i in document j, $N(w_i)$ is the number of documents containing term i, and R_k is the number of medline documents relevant to gene k. Cooley suggests that the specific nature of term weighting may not not be crucial to the performance of SVMs in text classification [11]

2.2 Classification

The assignment of yeast proteins to sub-cellular compartments was obtained from the MIPS web site [d]. According to MIPS, 2233 proteins have known locations in one or more of 16 categories. We limit our test and training data to these proteins. The locations and numbers of proteins at each location is shown in 1. For each location class, our training set consisted of half the number of genes that fall into this category plus half the of the remaining

[c] http://genome-www.stanford.edu/Saccharomyces/registry.html
[d] available from http://mips.gsf.de/proj/yeast/catalogues/subcell/index.html

negative examples. The test set consists of the remaining proteins - positive and negative cases.

Table 1: Number of positive examples in training and test sets for sub-cellular location

Role/location	+ve in training set	+ve in test set
organisation of plasma membrane	67	63
organisation of cytoplasm	279	245
organisation of cytoskeleton	47	52
organisation of endoplasmatic reticulum	68	80
organisation of Golgi	44	33
nuclear organisation	267	341
organisation of chromosome structure	19	18
mitochondrial organisation	174	155
peroxisomal organisation	19	12
vacuolar and lysosomal organisation	27	16
extracellular/secretion proteins	10	5

2.3 Training of SVM's

We used the support vector machine program SVM Light package v3.50 [14] [e]. We trained a SVM for each classification using a linear kernel function with C calculated as $\frac{1}{mean(x \cdot x)}$.

2.4 Evaluation

We evaluate the classification performance using a variety of methods. For traditional text retrieval evaluation measures based on precision/recall have been widely used [15]. We use precision/recall plots calculated on the distance of each test vector from the SVM decision boundary; however, comparison of performance between them is difficult because the classes contain different numbers of positive examples. To assess the global performance of classification methods we employed micro- and macro- averaging of the precision/recall data. Micro-averaging determines precision and recall of a set of binary classifiers averaged over the number of documents; this equates to evaluating average

[e] available from http://ais.gmd.de/thorsten/svm_light

performance a document selected randomly from the test collection. In macro-averaging, the recall/precision are averaged over the number of classes. Macro-averaging estimates the expected performance of an SVM trained on a new class; whereas micro-averaging estimates the performance of the system with new documents. For our purposes, micro-averaging is more useful.

We also use the F1 measure proposed by van Rijsbergen [7]. F1 is given by $\frac{2rp}{r+p}$ where p and r are precision and recall respectively. We determine the maximal value of F1 for the performance of each system on a particular classification.

3 Results

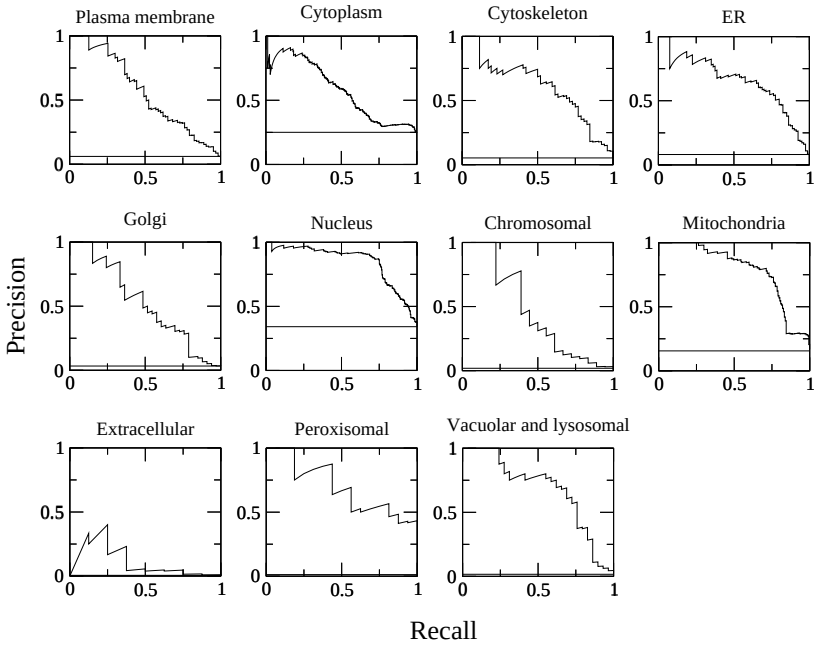

Figure 1: Precision/recall plots for location classifiers trained on term vectors. Horizontal lines indicate the performance of a random classifier

Precision/recall graphs for the various classifications are shown in figure 1. The performance of a random classifier is shown as a horizontal line in each plot. At low levels of recall, the precision is generally very high (95%+).

Classes with a large number of positive examples - nuclear, cytoplasmic, and mitochondrial - are better predicted than the rarer classifications. This is reflected in averaged precision/recall shown in figure 2. The better apparent performance from micro-averaging is a result of better prediction of bigger classes. The values of $Max(F1)$ are shown in table 2 and are much greater in all cases than a random classifier.

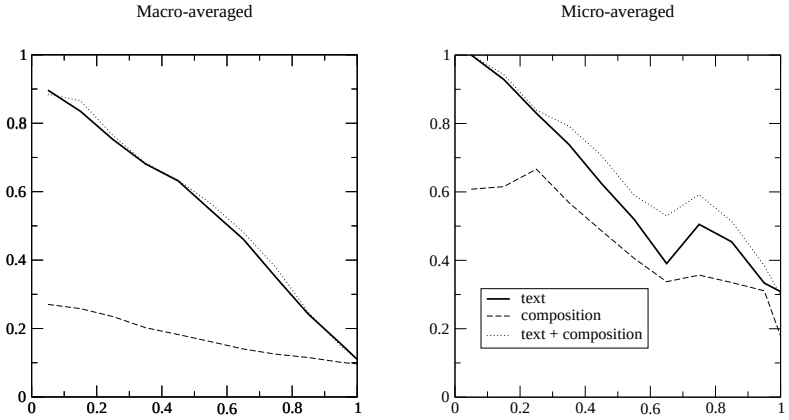

Figure 2: Micro and macro averaging of classification to 11 locational categories.

3.1 Sequence and text together improve classification

It has been known for some time that the amino acid composition of a protein can be used as an indicator of its sub-cellular localisation [16,2]. In particular, membrane associated proteins tend towards hydrophobicity, while intracellular proteins tend to be low in cysteine and rich in aliphatic and charged amino acids. Nuclear proteins generally contain disproportionally more charged and polar residues.

Figure 3 illustrates the performance of support vector machines in discriminating protein localisation based of their fractional composition of the twenty amino acids. It can been seen that composition is a poor predictor of ER, cytoskeleton, golgi, peroxisomal and vacuolar proteins, but good at

Table 2: Maximum F1-value for classifications

Role/location	Max F1			
	text alone	text + composition	composition alone	random
organisation of plasma membrane	0.54	0.56	0.47	0.12
organisation of cytoplasm	0.55	0.60	0.48	0.39
organisation of cytoskeleton	0.62	0.61	0.13	0.10
organisation of endoplasmatic reticulum	0.65	0.66	0.10	0.14
organisation of Golgi	0.54	0.53	0.10	0.14
nuclear organisation	0.80	0.82	0.61	0.51
organisation of chromosome structure	0.52	0.52	0.20	0.04
mitochondrial organisation	0.75	0.75	0.36	0.27
peroxisomal organisation	0.67	0.65	0.03	0.01
vacuolar and lysosomal organisation	0.69	0.69	0.06	0.02
extracellular/secretion proteins	0.31	0.33	0.12	0.01

predicting cytoplasmic, membrane and nuclear proteins. Composition also contains limited information on mitochondrial and chromosomal proteins. For extra-cellular proteins the scarcity of data makes assessment difficult, but the composition of these proteins gives better than random predictions.

When the test and training vectors derived from terms within Medline are extended to include the amino acid composition, performance in classifying proteins to the cytosol and nucleus is enhanced (table 2). In particular recall is improved. This may reflect improved performance on those proteins which have relatively few citations in the literature.

3.2 Detecting errors in annotation

Any manual method of gene annotation is liable to errors of omission and mis-classification. We checked apparent false negatives and positives to assess whether they were genuine by inspection of the relevant Medline documents.

For the cytosolic classification, the top scoring 'false' positive is cdc42, a Rho-type GTPase involved in bud site assembly and cell polarity. cdc42 contains a CAAX motif for geranylgeranyl modification and is likely to be associated with cell membranes. Ziman *et al.*,[17] determined that cdc42 exists in both a soluble form and membrane associated form within the cell; thus cdc42 should be included in cytosolic classification. A similar situation exists with ypt1 which is a GTP-binding protein required for vesicle transport from ER to Golgi and within the Golgi stack. It also undergoes geranylgeranyl modification, but the abundance and significance of any cytosolic form of the protein is not clear.

There are several proteins which MIPS assigns to the nucleus that our

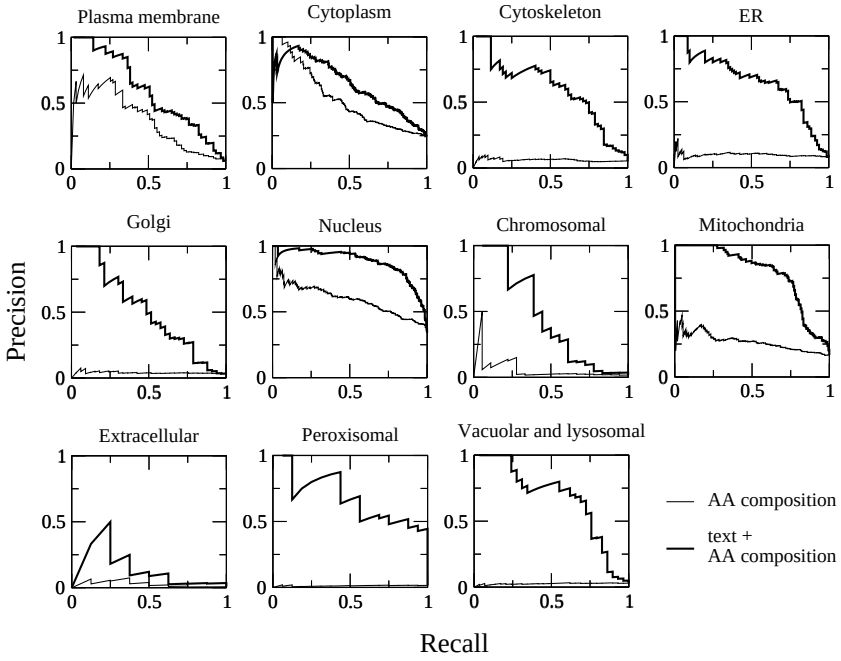

Figure 3: Precision/recall plots for location classifiers trained on term and/or amino acid composition vectors.

method correctly flags as being non-nuclear. These include: UBC6 - a ubiquitin-conjugating enzyme, anchored in the ER membrane with the catalytically active domain in cytoplasm [18]; hts1 - a histidyl-tRNA synthetase which is located exclusively in the mitochondria and cytosol [19]; and SMI1 protein involved in beta-1,3-glucan synthesis which has been shown to localise in patches at bud sites [20].

4 Discussion

4.1 Functional classification using SVMs and text

Here we show that functional classification of genes can be facilitated by text analysis of documents relevant to a gene. Other than a list of gene naming terms and synonyms, our method uses no prior knowledge of the problem domain nor any information from previously compiled sequence databases.

Automatic functional assignment of proteins can be used to improve manual assignment by spotting errors and increasing recall. Such errors may be simple mistakes, or the result of partial or incorrect information or understanding on the part of the human classifier. Even in the absence of such errors, assessments of what constitutes a correct assignment of documents into a classification will vary from user to user; thus there is a theoretical limit to the precision of an automatic classifier. For nuclear and mitochondrial proteins, automatic classification may be approaching this limit.

Although amino acid composition is generally a poor indicator of subcellular location, for some locations sequence provides a strong signal. In such cases, combining text and composition features can enhance recall.

4.2 Comparison with other methods

To compare our classification methods to that of Eisenhaber and Bork, we tested their algorithm (Meta-Annotator) on a subset of our original data that is present in SWISS-PROT [4]. Meta-Annotator is outstandingly good at predicting mitochondrial proteins and very good at predicting nuclear proteins.

Because Meta-Annotator joins the golgi and endoplasmic reticulum (ER) into a single class, we modified our treatment of this locational class. A single SVM trained to distinguish golgi or ER from others performed very poorly, probably because the intersection of these two sets is very small (8 cases) according to the MIPS classification. We therefore used the max(F1) value from micro-averaging of two SVMs trained on the ER and golgi proteins independently. Text classification using SVMs out-performs Meta-Annotator for cytoplasmic and golgi/ER proteins.

It should be borne in mind when comparing the two approaches that Meta-Annotator involves a large ammount of manual intervention. Not only is the method only applicable to a previously manually curated protein database (Swiss-Prot), but it also has encoded into more than 1000 logical rules derived from a human expert. Our approach requires no human input other than a list of gene names and synonyms. Given these facts it is little wonder than Meta-Annotator can generally out-perform our method. It is encouraging that a generic automatic approach can perform so well. With a larger set of training documents the SVM approach may be improved.

4.3 Combining features for functional classification of proteins

In this paper we have demonstrated that combining disparate features of a protein can aid in the functional classification of that protein. With the advent of many high-throughput studies of genes and proteins, many more features can

Table 3: Comparison of text SVMs and Meta-Annotator

Role/location	Meta_A precsion/recall	Meta$_A F1$	max(F1) for text
organisation of cytoplasm	49/32	0.38	0.54
organisation of Golgi/ER	75/48	0.58	0.62
nuclear organisation	87/86	0.86	0.80
mitochondrial organisation	90/93	0.91	0.75

be used as training data for binary classifiers. These include protein interaction data, features of the protein or DNA sequence and expression array data. The inclusion of a variety of independent or semi-independent features should improve recall since data for every protein may not be available from every experiment. For example, our method can be applied to proteins/genes of unknown sequence or conversely, sequence information can be used to infer function in the absence of any text relevant to the protein/gene.

Support vector machines are well suited to classification tasks of high dimensionality in which many features may be noisy or irrelevant. There is no doubt that data from expression array and protein interaction experiments can yield insights into gene function, but the quality of such data is hard to determine. The method presented here may ameliorate some of these problems.

Finally, entities other than proteins and genes can be represented as high dimensional vectors of text terms; these include whole organisms, protein complexes, protein domains or motifs, small molecules, cells and arbitrary text documents. In short, any entity which contains text, or for which relevant texts can be retrieved can be placed within a classification scheme.

5 References

1. F. Eisenhaber and P. Bork. Wanted: subcellular localization of proteins based on sequence. *Trends Cell Biol*, 8(4):169–70, Apr 1998.
2. K. Nishikawa and T. Ooi. Correlation of the amino acid composition of a protein to its structural and biological characters. *J Biochem (Tokyo)*, 91(5):1821–4, May 1982.
3. F. Eisenhaber and P. Bork. Evaluation of human-readable annotation in biomolecular sequence databases with biological rule libraries. *Bioinformatics*, 15(7-8):528–35, Jul-Aug 1999.
4. A. Bairoch and R. Apweiler. The protein sequence data bank and its supplement trembl in 1999. *Nucleic Acids Res*, 27(1):49–54, Jan 1 1999.
5. L. Wong. PIES, a protein interaction extraction system. In *Pac Symp Biocomput*, pages 520–31, Hawaii, 2001.

6. S. K. Ng and M. Wong. Toward routine automatic pathway discovery from on-line scientific text abstracts. In *Genome Inform Ser Workshop Genome*, volume 10, pages 104–112., 1999.
7. T. C. Rindflesch, L. Tanabe, J. N. Weinstein, and L. Hunter. Edgar: extraction of drugs, genes and relations from the biomedical literature. In *Pac Symp Biocomput*, pages 517–28., Hawaii, 2000.
8. C. J. van Rijsbergen. *Information Retrieval.* Butterworths, 2nd edition, 1979.
9. V.N. Vapnik. *The nature of statistical learning theory.* Springer Verlag, Heidelberg, DE, 1995.
10. V. Vapnik, S. Golowich, and A. Smola. Support vector method for function approximation, regression estimation, and signal processing. In M.C. Mozer, M.I. Jordan, and T. Petsche, editors, *Advances in Neural Information Processing Systems*, volume 9, pages 281–287, 1997.
11. R. Cooley. Classication of news stories using support vector machines. In *International Joint Conference on Articial Intelligence Text Mining Workshop*, 1999.
12. J. T-Y. Kwok. Automated text categorization using support vector machine. In *Proceedings of the International Conference on Neural Information Processing (ICONIP)*, pages 347–351, Kitakyushu, Japan, 1999.
13. J. M. Cherry, C. Ball, K. Dolinski, S. Dwight, M. Harris, J. C. Matese, G. Sherlock, G. Binkleyand H. Jin, S. Weng, and D. Botstein. Saccharomyces genome database. ftp://genome-ftp.stanford.edu/pub/yeast/SacchDB/, 2000.
14. T. Joachims. Text categorization with support vector machines: learning with many relevant features. In *10th European Conference on Machine Learning*, pages 137–142, Chemnitz, DE, 1998. Springer Verlag, Heidelberg, DE.
15. R. Baeza-Yates and B. Ribeiro-Neto. *Modern information retrieval.* Addison-Wesley, Harlow, England, 1999.
16. J. Cedano, P. Aloy, J. A. Perez-Pons, and E. Querol. Relation between amino acid composition and cellular location of proteins. *J Mol Biol*, 266(3):594–600., Feb 28 1997.
17. M. Ziman, D. Preuss, J. Mulholland O., J. M. 'Brien, D. Botstein, and D. I. Johnson. Subcellular localization of cdc42p, a saccharomyces cerevisiae-binding protein involved in the control of cell polarity. *Mol Biol Cell*, 4(12):1307–16., Dec 1993.
18. U. Lenk and T. Sommer. Ubiquitin-mediated proteolysis of a short-lived regulatory protein depends on its cellular localization. *J Biol Chem*, 275(50):39403–10., Dec 15 2000.

19. M. I. Chiu, T. L. Mason, and G. R. Fink. Hts1 encodes both the cytoplasmic and mitochondrial histidyl-trna synthetase of saccharomyces cerevisiae: mutations alter the specificity of compartmentation. *Genetics*, 132(4):987–1001., Dec 1992.
20. H. Martin, A. Dagkessamanskaia, G. Satchanska, N. Dallies, and J. Francois. Knr4, a suppressor of saccharomyces cerevisiae cwh mutants, is involved in the transcriptional control of chitin synthase genes. *Microbiology*, 145:249–58, Jan 1999.