BIOINFORMATICS

Vol. 1 no. 1 2003
Pages 1–8

# ***Predicting gene function in*** *Saccharomyces cerevisiae*

*A. Clare and R. D. King*

*Department of Computer Science, University of Wales, Aberystwyth, Penglais, Aberystwyth, SY23 3DB, Wales*

**ABSTRACT**

**Motivation:** *S. cerevisiae* is one of the most important model organisms, and has has been the focus of over a century of study. In spite of these efforts, 40% of its open reading frames (ORFs) remain classified as having unknown function (MIPS: Munich Information Center for Protein Sequences). We wished to make predictions for the function of these ORFs using data mining, as we have previously successfully done for the genomes of *M. tuberculosis* and *E. coli*. Applying this approach to the larger and eukaryotic *S. cerevisiae* genome involves modifying the machine learning and data mining algorithms, as this is a larger organism with more data available, and a more challenging functional classification.

**Results:** Novel extensions to the machine learning and data mining algorithms have been devised in order to deal with the challenges. Accurate rules have been learned and predictions have been made for many of the ORFs whose function is currently unknown. The rules are informative, agree with known biology and allow for scientific discovery.

**Availability:** All predictions are freely available from http://www.genepredictions.org, all datasets used in this study are freely available from http://www.aber.ac.uk/compsci/Research/bio/dss/yeastdata and software for relational data mining is available from http://www.aber.ac.uk/compsci/Research/bio/dss/polyfarm.

**Key words:** yeast, *S. cerevisiae*, DMP, prediction, functional genomics, scientific discovery

**Contact:** afc@aber.ac.uk

## INTRODUCTION

*Saccharomyces cerevisiae* (baker's or brewer's yeast) is a model eukaryotic organism. It was the first eukaryotic genome sequence to be completed (Goffeau *et al.*, 1996). It is cheap and quick to grow, generally non-pathogenic, and easy to manipulate genetically, and it has been the focus of detailed study over the years. These studies include expression analysis via Northern blots (Richard *et al.*, 1997), SAGE (Velculescu *et al.*, 1997) and microarrays (DeRisi *et al.*, 1997; Cho *et al.*, 1998; Chu *et al.*, 1998; Spellman *et al.*, 1998; Gasch *et al.*, 2000), 2-d gel electrophoresis (Boucherie *et al.*, 1995), two-hybrid systems for protein-protein interactions (Fromont-Racine *et al.*, 1997), large scale deletion and mutational analysis (Ross-Macdonald, 1999; Kumar *et al.*, 2000; Oliver, 1996) and phenotypic analysis (Oliver, 1996).

Despite all this biological knowledge and investigation, approximately 40% of the potential genes or Open Reading Frames (ORFs) in yeast remain without clear function or purpose.

Many computational methods are used in functional genomics. The most common method of determining the function of a protein is to use a sequence similarity program such as PSI-BLAST (Altschul *et al.*, 1997) to infer function by orthologous homology.

One method that has been shown to work well in predicting function is Data Mining Prediction (DMP) (King *et al.*, 2000). This method uses data mining and machine learning to induce rules predicting function from a variety of data sources, and is more general than conventional sequence similarity methods. Recently, we have successfully applied DMP to both the *M. tuberculosis* and *E. coli* genomes (King *et al.*, 2001), and made many predictions for their ORFs.

*S. cerevisiae* is larger and more complex than *M. tuberculosis* or *E. coli*. It has 16 chromosomes, which contain approximately 6,300 ORFs. More data is associated with each of these ORFs, and a wider variety of data sources are available. We used the MIPS† functional classification scheme for these ORFs. We preferred this to GO‡ because more ORFs were annotated with a known class through the MIPS scheme.

This paper applies machine learning and data mining to the problem of predicting ORF functional class in the genome of *S. cerevisiae*.

## DATA

A wide variety of data sources are available for yeast. We chose to use 5 different types of data.

### Sequence data

This dataset consisted of data that can be directly calculated from sequence, such as amino acid ratios, molecular weight, aliphatic index, codon adaption index, sequence length and hydrophobicity. The ProtParam tool§ was used

† http://mips.gsf.de/proj/yeast/CYGD/
‡ http://www.geneontology.org/
§ http://www.expasy.org/tools/protparam.html

to calculate many of these attributes. This dataset consisted of 478 attributes in total (most of them ratios of amino acid pairs).

### Phenotype data

This is data from phenotypic growth experiments, where knockout mutants are grown in a variety of media with the aim of finding growth conditions where the mutant and the wild type (no mutation) differ ("a phenotype"). Data was collected from three separate sources: TRIPLES (Kumar *et al.*, 2000), EUROFAN (Oliver, 1996) and MIPS (Mewes *et al.*, 1999). This dataset consisted of 69 attributes in total.

### Expression data

Many genome-wide microarray experiments have been carried out on yeast. We collected 7 sources of data: cellcycle (Spellman *et al.*, 1998), church (Roth *et al.*, 1998), derisi (DeRisi *et al.*, 1997), eisen (Eisen *et al.*, 1998), gasch1 (Gasch *et al.*, 2000), gasch2 (Gasch *et al.*, 2001), and spo (Chu *et al.*, 1998). These experiments measure changes in expression levels under a variety of conditions, such as heat shock, and during the normal lifetime of the cell.

### Homology data

Sequence similarity searches are a standard method of inferring function. Our homology data is the result of a PSI-BLAST search for each *S. cerevisiae* ORF against NRDB90¶. We also joined on to NRDB90 the yeast genome itself, so that similar sequences within the genome could be discovered.

The sequences that are found by PSI-BLAST to have an e-value below the threshold are known as "hits". For each ORF, we looked up certain data from the SWISSPROT entries that were hits for the ORF. We used SWISSPROT version 39. The data extracted from the SWISSPROT entries were organism classification, database cross references, and certain keywords.

This relational data was expressed in Datalog‖.

### Predicted secondary structure data

Data was collected about the predicted secondary structure of the gene products of *S. cerevisiae*. Prof (Ouali & King, 2000) was used to make the predictions. The predictions were expressed as Datalog facts representing the lengths and relative positions of the alpha, beta and coil parts of the structure. The predictions also included the distributions of alpha, beta and coil as percentages.

The homology and secondary structure datasets are relational in nature, whereas the other data sets are straightforward attribute-value data. All data sets used are shown in Table 1.

| Name | Description |
|---|---|
| seq | Data consisting only of attributes that can be calculated from sequence alone (for example amino acid ratios, sequence length and molecular weight) |
| pheno | Data from phenotype growth experiments |
| struc | Data from secondary structure prediction. Boolean attributes were constructed from the first order patterns mined by PolyFARM. |
| hom | Data from the results of PSI-BLAST searches of NRPROT. Boolean attributes were constructed from the first order patterns mined by PolyFARM. |
| cellcycle | Microarray data from (Spellman *et al.*, 1998) |
| church | Microarray data from (Roth *et al.*, 1998) |
| derisi | Microarray data from (DeRisi *et al.*, 1997) |
| eisen | Microarray data from (Eisen *et al.*, 1998) |
| gasch1 | Microarray data from (Gasch *et al.*, 2000) |
| gasch2 | Microarray data from (Gasch *et al.*, 2001) |
| spo | Microarray data from (Chu *et al.*, 1998) |
| expr | All microarray datasets concatenated together. |

**Table 1.** Datasets used in this work

### Function classification

The MIPS classification scheme is hierarchical with 19 general classes at the top level that are subdivided into more specific classes at the second level. These are in turn subdivided, and then again, so the hierarchy is four levels deep. An ORF can (and usually does) belong to several classes in various positions in this hierarchy, as the proteins can be involved in more than one function within the cell.

## METHODS

The use of machine learning and data mining for making predictions for ORF function has been previously successfully demonstrated on the genomes of *M. tuberculosis* and *E. coli* (King *et al.*, 2001). We call this method DMP (Data Mining Prediction), and two complementary forms of data mining are used: Inductive Logic Programming (Muggleton, 1992; Lavrač & Džeroski, 1994) and propositional rule learning (Mitchell, 1997).

When DMP was applied to the *M. tuberculosis* and *E. coli* data, the Inductive Logic Programming system WARMR (Dehaspe & De Raedt, 1997) was used to extract interesting patterns from the relational datasets. These

¶NRDB90 is a non-redundant protein database where proteins that share more than 90% similarity have been removed (Holm & Sander, 1998). This is created from the union of the SWISSPROT, SWISSNEW, TREMBL, TREMBLNEW, GenBank, PIR, WormPep and PDB databases. We used the version as of 4th January 2001 from http://www.ebi.ac.uk/~holm/nrdb90/, containing 260,000 non-duplicate sequences. PSI-BLAST was used with the following parameters: "-e 10 -h 0.0005 -j 20". The version of PSI-BLAST was BLASTP 2.0.12 [Apr-21-2000].

‖Datalog is the language of function free and negation free Horn clauses (Prolog without functions). As a database query language it has been extensively studied (Ullman, 1988).

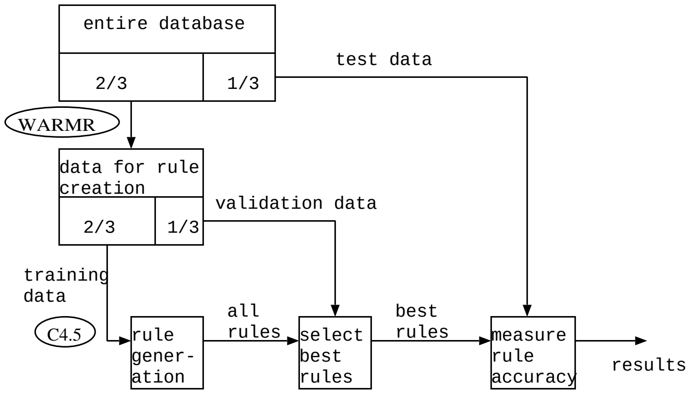


**Fig. 1.** The structure of the DMP method. 1/3 of the data is held out for later testing the accuracy. WARMR is used to extract patterns from relational data. These patterns are used as attributes and combined with the non-relational data. 2/3 of this is used as training data for C4.5, the remaining 1/3 is held out as validation data. The rules produced are filtered by comparison with the validation data, and then the accuracy of the filtered ruleset is measured on the held out test data.

patterns were used as boolean attributes (1 indicates presence of this pattern for an ORF, and 0 indicates absence). The propositional rule learner C4.5 (Quinlan, 1993) was applied to this boolean attribute data together with the data from the other datasets, to produce classification rules. Statistically significant rules were selected on a held out validation dataset, and their accuracy measured on a held out test data set. Figure 1 shows the scientific methodology used.

For yeast, it was necessary to make modifications to the original machine learning and data mining methods. Yeast has a larger genome than *M. tuberculosis* or *E. coli*, it is eukaryotic (therefore more homologous proteins in the database), and more post-genomic experimental data is available. The WARMR algorithm that had been used previously was not empirically able to process the larger amount of relational data that we had for yeast. It was therefore necessary to develop a new algorithm (PolyFARM) based on WARMR, but specific to our task, and able to run in a distributed fashion on a Beowulf cluster, to make use of the processing power available. PolyFARM is the first system to the best of our knowledge for distributed first order association mining (Clare & King, 2003).

Another complication is that the functional classification scheme for yeast allowed ORFs to belong to more than one class. Such multi-label classification problems have been little studied in data mining, and so this required making changes to our base algorithm (C4.5) to deal with this (Clare & King, 2002). C4.5 is a decision tree algorithm which uses entropy as its internal measurement of the best decision to make at each stage. The entropy calculations were modified to measure the uncertainty in data represented by sets of classes rather than singleton classes. Also the program's data structures, counting methods, pruning methods and rule generation needed to be modified.

A further complication is that the functional classification scheme is a hierarchy, so the class labels were not necessarily independent of one another. There has been little prior work using hierarchical classes. Most work in this area has been done in relation to classifying large volumes of text documents. The classification algorithms used in these text processing applications tend to be either clustering or very simple statistical algorithms such as naïve Bayes, working on high volumes of data. Mitchell (1998) demonstrated that a hierarchical Bayesian classifier would have the same performance as a flat Bayesian classifier under certain assumptions: smoothing is not used to estimate the probabilities and the same features are used by different classifiers in the hierarchy. Work has also been done on smoothing probabilities of items in low frequency classes by making use of their parent frequencies (McCallum *et al.*, 1998) and making more specific classifiers using different features at different places in the hierarchy (Koller & Sahami, 1997; Chakrabarti *et al.*, 1998; Mladenic & Grobelnik, 1998).

Recently, Blockeel *et al.* (2002) have also designed an algorithm to tackle the problem of hierarchical multi-classification. They construct a type of decision tree called a "clustering tree" to do the classification, where the criteria for deciding on how to split a node is based on minimizing the intra-cluster variance of the data within a node. The distance measure used to calculate intra-cluster variance works on sets of class labels and takes into account both the class hierarchy and the multiple labels. They have applied their method to our phenotype data and obtained a small tree with just 2 tests.

In this work we continued to use our multi-label modified version of C4.5, and modified it further to use a classification hierarchy. This involved changes to read and store the class hierarchy, test for membership of a class, find the best class or classes to represent a node, and perform the entropy calculations. Membership of a class implies membership of all its parent classes. Representing a node by classes from the most general level of the hierarchy will give a lower entropy than using more specific classes. However, we prefer more specific classes for the results, since this gives more biological knowledge. The entropy formula was thus weighted to account for this.

The results from this hierarchical version of C4.5 were to be compared with the results of treating the hierarchy level by level, as independent classes.

Another form of hierarchy comes from the organism classification data from SWISSPROT used in our homology dataset. This is part of a taxonomy. The homology dataset is relational, so was processed with PolyFARM. The original program WARMR that was used to process the *M. tuberculosis* and *E. coli* data used recursive predicates to define the hierarchy. PolyFARM defines the hierarchy more directly as a tree structure, and propagates association counts back up the tree to parent nodes. The resulting associations do make use of terms at various levels in this taxonomy.

## RESULTS

### Validation

All results tables are given for the rulesets after validation has been applied (i.e. just the significant rules). The validation was applied by keeping only the rules which were shown to be statistically significant on the validation data set. Statistical significance was calculated by using the hypergeometric distribution with an $\alpha$ value of 0.05 and a Bonferroni correction.

### Summary

Table 2 shows the testset average accuracy of each of the rulesets produced from the different types of data. We list the accuracies produced from learning on each level of the classification hierarchy in turn with the multilabel version of C4.5 (level 1 is the most general and level 4 the most specific functional annotation). We also list for comparison the accuracy of the hierarchical version of C4.5, which makes use of multilabel classes at all levels.

The accuracies range between 75% and 39% on level 1, dropping on lower levels where the data is sparse. These accuracies are highly significant when compared to the *a-priori* class probabilities. To demonstrate this, the *a-priori* probabilities for level 1 classes can be seen in Table 3. This table also gives a class by class breakdown of the accuracies, revealing that some classes are consistently predicted much better than others, and some types of data predict certain classes better than others. Class 5,0,0,0 - "protein synthesis" (and in particular its subclass 5,1,0,0 - "ribosome biogenesis") is consistently predicted very well by most datasets, especially the expression datasets. Its *a priori* probability is just 9% but the rulesets are between 55% and 93% accurate on this class.

The accuracy of our version of C4.5 that was modified to make use of the class hierarchy is sometimes better than that of the non-hierarchical C4.5 on the individual levels, and sometimes worse (Table 2). This is disappointing as we would expect that given more information the results should be consistently better. Also the rules that were produced by the hierarchical C4.5 were different to those produced on the individual levels. This is to be expected as the criteria for choosing nodes to construct the decision tree are slightly different, and a different amount of information is available.

| | level | | | | |
|---|---|---|---|---|---|
| datatype | 1 | 2 | 3 | 4 | all |
| seq | 55 | 55 | 33 | 0 | 71 |
| pheno | 75 | 40 | 7 | 0 | 68 |
| struc | 49 | 43 | 0 | 0 | 58 |
| hom | 65 | 38 | 69 | 20 | 55 |
| cellcycle | 63 | 33 | 21 | 0 | 54 |
| church | 75 | 43 | 0 | 0 | 53 |
| derisi | 64 | 51 | 0 | 0 | 61 |
| eisen | 63 | 40 | 28 | 0 | 48 |
| gasch1 | 39 | 46 | 44 | 75 | 38 |
| gasch2 | 44 | 66 | 40 | 0 | 60 |
| spo | 43 | 63 | 0 | 0 | 46 |
| expr | 42 | 37 | 35 | 0 | 75 |

**Table 2. Accuracy**: Average accuracy (percentages) on the test data of each ruleset. Only rules which were statistically significant on the validation set are included. Level "all" indicates the results of the hierarchical version of C4.5, which had classes from all levels in its training data. Significance was calculated by the hypergeometric distribution, with alpha=0.05 and Bonferroni correction.

| | level | | | | |
|---|---|---|---|---|---|
| datatype | 1 | 2 | 3 | 4 | all |
| seq | 2240 (1646) | 39 (39) | 38 (38) | 0 (0) | 156 (147) |
| pheno | 25 (25) | 64 (64) | 44 (44) | 0 (0) | 0 (0) |
| struc | 114 (114) | 109 (99) | 0 (0) | 0 (0) | 29 (27) |
| hom | 133 (82) | 325 (301) | 4 (4) | 13 (13) | 49 (48) |
| cellcycle | 993 (961) | 785 (748) | 392 (392) | 0 (0) | 1910 (1544) |
| church | 4 (4) | 75 (49) | 0 (0) | 0 (0) | 1333 (1079) |
| derisi | 1164 (1144) | 148 (129) | 0 (0) | 0 (0) | 74 (59) |
| eisen | 9 (9) | 32 (24) | 15 (15) | 0 (0) | 15 (13) |
| gasch1 | 918 (873) | 737 (714) | 35 (35) | 21 (21) | 1232 (1065) |
| gasch2 | 203 (201) | 212 (194) | 12 (12) | 0 (0) | 1732 (1522) |
| spo | 174 (174) | 116 (104) | 0 (0) | 0 (0) | 221 (210) |
| expr | 1133 (1066) | 1416 (1175) | 150 (149) | 0 (0) | 52 (42) |

**Table 4. Predictions:** Predictions for ORFs of unknown function (classes 99,0,0,0 and 98,0,0,0) made by the rulesets. Numbers of predictions made are given with actual numbers of ORFs in brackets, as there may be more than one class predicted for each ORF. Only rules which were statistically significant on the validation set are used. Significance was calculated by the hypergeometric distribution, with alpha=0.05 and Bonferroni correction.

The rules were then applied to the ORFs whose function is currently unknown. The numbers of predictions made for ORFs whose function is currently unknown can be found in Table 4. The **seq** data makes the most predictions at level 1 with 1646 ORFs assigned some function. However this drops sharply, with **seq** predicting only 39 ORFs at level 2. The expression data sets also make large numbers of predictions, and the combined expression data makes a huge number of predictions.

| class | prior | seq | pheno | struc | hom | cellcycle | church | derisi | eisen | gasch1 | gasch2 | spo | expr |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| metabolism | 27 | 52 | | | 65 | | | | 53 | 41 | | | 47 |
| energy | 6 | | | | 20 | | | | 47 | 29 | | | 38 |
| cell cycle and DNA processing | 16 | | 75 | | | 67 | | | | 30 | 24 | 28 | |
| transcription | 20 | 33 | | | | | | | | | | | 33 |
| protein synthesis | 9 | 72 | | | 86 | 56 | 75 | 71 | 91 | 81 | 74 | 81 | 93 |
| protein fate | 15 | | | | 100 | | | | | 35 | | | |
| cellular transport | 12 | | | | 42 | | | | | | | | |
| transposable elements, viral and plasmid proteins | 3 | 85 | | 33 | 63 | | | | | | | | |
| control of cellular organization | 5 | | 75 | | | | | | | | | | |
| subcellular localization | 57 | 61 | 75 | | | 64 | | 64 | | | | | |
| transport facilitation | 8 | 42 | | 49 | 77 | | | | | 24 | | | |

**Table 3.** Class by class accuracies (percentages) for rulesets produced by the datasets, level 1 in the functional classification hierarchy. Blank entries indicate that there were no rules produced for this clsss from this dataset.

| | |
|---|---|
| **if** | no: coil followed by beta followed by coil (c-b-c) |
| **and** | yes: coil (of length 3) followed by alpha ($10 \leq$ length $< 14$) |
| **and** | yes: coil (of length 1 or 2) followed by alpha ($10 \leq$ length $< 14$) |
| **and** | yes: coil (of length 3) followed by alpha ($3 \leq$ length $< 6$) |
| **and** | no: coil ($6 \leq$ length $< 10$) followed by alpha (of length 1 or 2) |
| **then** | the function of this ORF is 8/4/0/0 "mitochondrial transport" |

**Fig. 2.** Rule 76, level 2, **structure** data

## Biological understanding

The rules are informative and can be used to understand more about the biology behind the predictions. Here we demonstrate that the rules can be shown to be consistent with known biology.

*Rule from structure data* Figure 2 shows a rule from level 2 of the structure data. This rule is 80% accurate on the test set.

The rule requires several predicted short coils followed by fairly long alphas, and this happens at least 3 times. There is no predicted pattern of coil-beta-coil, and there are no predicted fairly long coils followed by short alphas.

Most of the ORFs that match this rule belong to the Mitochondrial Carrier Family (MCF). These are known to have six transmembrane a-helical spanning regions (Kuan & Saier Jr., 1993). Kuan and Saier produced a multiple alignment of members of the MCF and analysed hydropathy plots. They observed that "These analyses revealed that the six transmembrane spanners exhibited varying degrees of sequence conservation and hydrophilicity. These spanners, and immediately adjacent hydrophilic loop regions, were more highly conserved than other regions of these proteins."

The alpha-helices in these proteins are known to be long, in the order of 20-30 amino acids (Senes *et al.*, 2000). So we were curious to understand why this rule selects alpha stretches of only 10 to 14 amino acids. Using Clustal W (Higgins *et al.*, 1994) for a multiple alignment of the sequences of all the ORFs covered by this rule showed a few consensus positions, but nothing immediately obvious. Overlaying the secondary structure predictions given by Prof onto the alignment showed a striking pattern. Each long alpha helix was broken in the middle by one or two short predicted coils of 2-3 amino acids in length. All these short coils aligned perfectly and appeared at glycines and prolines in the sequences. Glycine is the smallest amino acid and may disrupt the helix. Proline is also known to cause kinks in helices, since it has an imino rather than an amino group. The rule detects these "kinks" in the helices.

In the multiple sequence alignment of all ORFs that fit this rule, helices 1, 3 and 5 have a conserved proline, whereas helices 2, 4 and 6 have a conserved GxxxxxxG motif. This motif is known to be associated with transporter/channel like membrane proteins (Liu *et al.*, 2002).

It is noteworthy that the Prof prediction rules were learnt from globular proteins, and there was no strong expectation that they would be of value for transmembrane sequences. These and other empirical results demonstrate the value of the information from protein secondary structure prediction proteins in analysing transmembrane proteins.

Errors of commission of this rule are:

```
YMR288W (HSH155)
''component of a multiprotein splicing factor''
YHR190W (ERG9)
''lipid, fatty-acid and isoprenoid biosynthesis,
  endoplasmic reticulum, farnesyl-diphosphate
```

| | |
|---|---|
| **if** | the ORF is not weakly homologous to a protein in klebsiella, |
| **and** | is strongly homologous to a protein in desulfurococcales, |
| **and** | is strongly homologous to a short protein in cyprinidae, |
| **then** | the function of this ORF is |
| | 6/0/0/0 "PROTEIN FATE (folding, modification, destination)" |

**Fig. 3.** Rule 1210, level 1, **homology** data

```
farnesyltransferase''
```

These are probably not members of the mitochondrial transport class. The rule is correct for the MCF members, but only recognises 3 parts of the alpha-helices, not all 6, and it is possible that there are other unrelated proteins that share this much of the structure.

This rule predicts two ORFs that are currently listed as being "unclassified proteins", YPR128C and YMR192W. One of these predictions has been shown to be correct. YPR128C has recently been shown to be a member of the mitochondrial carrier family (van Roermund *et al.*, 2001), although MIPS still does not make this classification. YMR192W, on the other hand, does not align well with the other sequences, either by primary or secondary structure, and we doubt that this prediction is correct.

*Rule from homology data* Figure 3 shows a rule from the most general classification level from the homology data. This rule is 100% accurate and remarkably simple.

On further investigation it can be seen that almost all ORFs that match this rule are from the 20S proteasome subunit. This subunit is part of the proteasome for degradation of proteins. Why should these particular 3 homologies be enough to specify this role?

- klebsiella is a eubacterium, a sister to escherichia, shigella and salmonella, found in our gastrointestinal tracts.
- desulfurococcales is an archaean, an anaerobic hyper-thermophile.
- cyprinidae is a eukaryote, a class of ray finned fish.

In eukaryotes the proteasome is the central enzyme of nonlysosomal protein degradation. It was originally thought that 20S protease subunits only existed in eukaryotes, but now they have also been found in all archaeal genomes sequenced. The actinomycetes phylum of bacteria is the **only** group of bacteria known to contain genuine 20S proteasomes (this branch includes the mycobacteria such as *M. tuberculosis*) (De Mot *et al.*, 1999; Zühl *et al.*, 1997). It would seem that they only exist in Gram positive bacteria with high GC content.

So this rule, which states that we need a strong homology to a eukaryotic protein, and an archaean protein, but not even a weak homology to a protein in the klebsiella bacterium, fits the current biological knowledge. It is perhaps surprising how few other (non-proteasomal) ORFs also fit this rule, and how unique this class of ORFs is. The need to ensure this is probably behind the strange choice of genera, e.g. cyprinidae.

## CONCLUSIONS

Many accurate and informative rules have been learnt that make predictions for the functions of ORFs of unknown function in *S. cerevisiae*. All the predictions can be obtained and investigated at http://www.genepredictions.org. This is a web database designed to make these predictions, and others, accessible to the bioinformatics community. This database aims to allow ORF function predictions to be widely available, while at the same time removes the problem of having unconfirmed annotations polluting the integrity of the main bioinformatic databases. The database can be searched either by ORF name or by functional class and presents results either as web documents or in spreadsheet compatible file format. Confidence estimates and links to the evidence for the predictions are provided. The rules themselves can be examined at http://www.aber.ac.uk/compsci/Research/bio/dss/yeastpreds/, and the datasets used are available at http://www.aber.ac.uk/compsci/Research/bio/dss/yeastdata/.

The basic DMP method devised for the *M. tuberculosis* and *E. coli* genomes was adaptable to the larger more complex genome of yeast. Modifications to the machine learning algorithm C4.5 were necessary, and a new version of the data mining algorithm was written to scale to the size of data.

Other work has been done previously using supervised machine learning to predict protein function. Decision tree learning has been used to annotate the Keyword field of proteins in SWISS-PROT (Kretschmann *et al.*, 2001). Features such as the InterProScan signature, and the organism classification line were used to generate rules to predict the presence or absence of a keyword. Hvidsten *et al.* (2001) used supervised learning (rough set theory) to predict GeneOntology classification of proteins from fibroblast serum response microarray data. Brown *et al.* (2000) used suport vector machines to discriminate 6 functional classes for yeast proteins from microarry data. Probabilistic decision tree learning has been used more recently (Syed & Yona, 2003) to predict the enzyme classification of SWISS-PROT proteins from sequence data, predicted secondary structure and SWISS-PROT database data. Our work is similar to theirs, with the addition of the data mining step for the relational data, and without using a mixture of probabilistic decision trees. Interestingly, they also have the problem of allowing multiple values for attributes, but rather than using data mining as a preprocessing step they use weightings within the decision tree. However, our homology data is too complex for this solution. We not only have multiple homologous proteins for each ORF, but also each homologous protein has multiple associated attributes.

Data mining and machine learning are powerful tools for functional genomics and extraction of knowledge from biological data. We have demonstrated that the DMP method can be scaled up to a larger genome and still give good estimated results. We are currently applying this method to the even larger genome of *Arabidopsis thaliana*, which has four times as many ORFs as yeast, and are confident that many more accurate predictions of ORF function will be made.


## ACKNOWLEDGEMENTS

Amanda Clare was supported by MRC grant G78/6609.


## REFERENCES


Altschul, S. F., Madden, T. L., Schaffer, A. A., Zhang, J., Zhang, Z., Miller, W. & Lipman, D. J. (1997). Gapped BLAST and PSI-BLAST: a new generation of protein database search programs. *Nucl. Acids. Res*, **25**, 3389–3402.

Blockeel, H., Bruynooghe, M., Dzeroski, S., Ramon, J. & Struyf, J. (2002). Hierarchical multi-classification. In Dzeroski, S., De Raedt, L. & S., W., (eds.) *Proceedings of the Workshop on Multi-Relational Data Mining (MRDM-2002)*. pp. 21–35.

Boucherie, H., Dujardin, G., Kermorgant, M., Monribot, C., Slonimski, P. & Perrot, M. (1995). Two dimensional protein map of *saccharomyces cerevisiae*: construction of a gene-protein index. *Yeast*, **11**, 601–613.

Brown, M., Nobel Grundy, W., Lin, D., Cristianini, N., Walsh Sugnet, C., Furey, T., Ares Jr., M. & Haussler, D. (2000). Knowledge-based analysis of microarray gene expression data by using support vector machines. *Proc. Nat. Acad. Sci. USA*, **97(1)**, 262–267.

Chakrabarti, S., Dom, B., Agrawal, R. & Raghavan, P. (1998). Scalable feature selection, classification and signature generation for organising large text databases into hierarchical topic taxonomies. *VLDB Journal*, **7**, 163–178.

Cho, R., Campbell, M., Winzeler, E., Steinmetz, L., Conway, A., Wodicka, L., Wolfsberg, T., Gabrielian, A., Landsman, D., Lockhart, D. & Davis, R. (1998). A genome-wide transcription analysis of the mitotic cell cycle. *Molecular Cell*, **2**, 65–73.

Chu, S., DeRisi, J., Eisen, M., Mulholland, J., Botstein, D., Brown, P. & Herskowitz, I. (1998). The transcriptional program of sporulation in budding yeast. *Science*, **282**, 699–705.

Clare, A. & King, R. D. (2002). Machine learning of functional class from phenotype data. *Bioinformatics*, **18(1)**, 160–166.

Clare, A. & King, R. D. (2003). Data mining the yeast genome in a lazy functional language. In *Practical Aspects of Declarative Languages (PADL'03)*.

De Mot, R., Nagy, I., Walz, J. & Baumeister, W. (1999). Proteasomes and other self-compartmentalizing proteases in prokaryotes. *Trends in Microbiology*, **7**, 88–92.

Dehaspe, L. & De Raedt, L. (1997). Mining association rules in multiple relations. In *7th International Workshop on Inductive Logic Programming*.

DeRisi, J., Iyer, V. & Brown, P. (1997). Exploring the metabolic and genetic control of gene expression on a genomic scale. *Science*, **278**, 680–686.

Eisen, M., Spellman, P., Brown, P. & Botstein, D. (1998). Cluster analysis and display of genome-wide expression patterns. *Proc. Nat. Acad. Sci. USA*, **95**, 14863–14868.

Fromont-Racine, M., Rain, J. & Legrain, P. (1997). Toward a functional analysis of the yeast genome through exhaustive two-hybrid screens. *Nature Genetics*, **16**, 277–282.

Gasch, A., Huang, M., Metzner, S., Botstein, D., Elledge, S. & Brown, P. (2001). Genomic expression responses to DNA-damaging agents and the regulatory role of the yeast ATR homolog Mec1p. *Mol. Biol. Cell*, **12**, 2987–3000.

Gasch, A., Spellman, P., Kao, C., Carmel-Harel, O., Eisen, M., Storz, G., Botstein, D. & Brown, P. (2000). Genomic expression program in the response of yeast cells to environmental changes. *Mol. Biol. Cell.*, **11**, 4241–4257.

Goffeau, A., Barrell., B., Bussey, H., Davis, R., Dujon, B., Feldmann, H., Galibert, F., Hoheisel, J., Jacq, C., Johnston, M., Louis, E., Mewes, H., Murakami, Y., Philippsen, P., Tettelin, H. & Oliver, S. (1996). Life with 6000 genes. *Science*, **274**, 563–7.

Higgins, D. G., J.D. Thompson, J. D. & Gibson, T. J. (1994). CLUSTAL W: improving the sensitivity of progressive multiple sequence alignment through sequence weighting, position specific gap penalties and weight matrix choice. *Nucleic Acids Research*, **22**, 4673–4680.

Holm, L. & Sander, C. (1998). Removing near-neighbour redundancy from large protein sequence collections. *Bioinformatics*, **14**, 423–429.

Hvidsten, T., Komorowski, J., Sandvik, A. & Lægreid, A. (2001). Predicting gene function from gene expressions and ontologies. In *Pacific Symposium on Biocomputing*. pp. 299–310.

King, R., Karwath, A., Clare, A. & Dehaspe, L. (2000). Genome scale prediction of protein functional class from sequence using data mining. In *KDD 2000*.

King, R., Karwath, A., Clare, A. & Dehaspe, L. (2001). The utility of different representations of protein sequence for predicting functional class. *Bioinformatics*, **17(5)**, 445–454.

Koller, D. & Sahami, M. (1997). Hierarchically classifying documents using very few words. In *ICML 97*.

Kretschmann, E., Fleischmann, W. & Apweiler, R. (2001). Automatic rule generation for protein annotation with the C4.5 data mining algorithm applied on SWISS-PROT. *Bioinformatics*, **17**, 920–926.

Kuan, J. & Saier Jr., M. (1993). The mitochondrial carrier family of transport proteins: structural, functional, and evolutionary relationships. *Crit. Rev. Biochem. Mol. Biol.*, **28(3)**, 209–33.

Kumar, A., Cheung, K.-H., Ross-Macdonald, P., Coelho, P., Miller, P. & Snyder, M. (2000). TRIPLES: a database of gene function in S. cerevisiae. *Nucleic Acids Res.*, **28**, 81–84.

Lavrač, N. & Džeroski, S. (1994). Inductive Logic Programming: Techniques and Applications. Ellis Horwood, Hemel Hempstead, UK.

Liu, Y., Engelman, D. M. & Gerstein, M. (2002). Genomic analysis of membrane protein families: abundance and conserved motifs. *Genome Biology*, **3(10)**.

McCallum, A., Rosenfeld, R., Mitchell, T. & Ng, A. (1998). Improving text classification by shrinkage in a hierarchy of classes. In *ICML 98*.

Mewes, H., Heumann, K., Kaps, A., Mayer, K., Pfeiffer, F., Stocker, S. & Frishman, D. (1999). MIPS: a database for protein sequences and complete genomes. *Nucleic Acids Research*, **27**, 44–48.

Mitchell, T. (1997). Machine Learning. McGraw-Hill, Singapore.

Mitchell, T. (1998). Conditions for the equivalence of hierarchical and non-hierarchical bayesian classifiers. Technical note, Carnegie Mellon University.

Mladenic, D. & Grobelnik, M. (1998). Learning document classification from large text hierarchy. In *AAAI 98*.

Muggleton, S., (ed.) (1992). Inductive Logic Programming. Academic Press.

Oliver, S. (1996). A network approach to the systematic analysis of yeast gene function. *Trends in Genetics*, **12(7)**, 241–242.

Ouali, M. & King, R. (2000). Cascaded multiple classifiers for secondary structure prediction. *Protein Science*, **9(6)**, 1162–76.

Quinlan, J. R. (1993). C4.5: programs for Machine Learning. Morgan Kaufmann, San Mateo, California.

Richard, G., Fairhead, C. & Dujon, B. (1997). Complete transcriptional map of yeast chromosome XI in different life conditions. *Journal of Molecular Biology*, **268**, 303–321.

Ross-Macdonald, P. e. a. (1999). Large-scale analysis of the yeast genome by transposon tagging and gene disruption. *Nature*, **402**, 413–418.

Roth, F., Hughes, J., Estep, P. & Church, G. (1998). Fining DNA regulatory motifs within unaligned noncoding sequences clustered by whole-genome mRNA quantitation. *Nature Biotechnology*, **16**, 939–945.

Senes, A., Gerstein, M. & Engelman, D. M. (2000). Statistical analysis of amino acid patterns in transmembrane helices: The GxxxG motif occurs frequently and in association with $\beta$-branched residues at neighboring positions. *Journal of Molecular Biology*, **296**, 921–936.

Spellman, P., Sherlock, G., Zhang, M., Iyer, V., Anders, K., Eisen, M., Brown, P., Botstein, D. & Futcher, B. (1998). Comprehensive identification of cell cycle-regulated genes of the yeast *Saccharomyces cerevisiae* by microarray hybridization. *Molecular Biology of the Cell*, **9**, 3273–3297.

Syed, U. & Yona, G. (2003). Using a mixture of probabilistic decision trees for direct prediction of protein function. In *RECOMB '03*.

Ullman, J. D. (1988). Principles of Database and Knowledge-Base Systems, Vol. 1 and 2. Computer Science Press, Rockville, Md.

van Roermund, C. W. T., Drissen, R., van den Berg, M., Ijlst, L., Hettema, R. H., Tabak, H. F., Waterham, H. R. & Wanders, R. J. A. (2001). Identification of a peroxisomal ATP carrier required for medium-chain fatty acid $\beta$-oxidation and normal peroxisome proliferation in *Saccharomyces cerevisiae*. *Molecular and Cellular Biology*, **21(13)**, 4321–4329.

Velculescu, V., Zhang, L., Zhou, W., Vogelstein, J., Basrai, M., Bassett, D., Heiter, P., Vogelstein, B. & Kinzler, K. (1997). Characterization of the yeast transcriptome. *Cell*, **88**, 243–251.

Zühl, F., Tamura, T., Dolenc, I., Cejka, Z., Nagy, I., De Mot, R. & Baumeister, W. (1997). Subunit topology of the *Rhodococcus* proteasome. *FEBS Letters*, **400**, 83–90.