Data and text mining

Functional bioinformatics for Arabidopsis thaliana

A. Clare\textsuperscript{1,*}, A. Karwath\textsuperscript{2}, H. Ougham\textsuperscript{1} and R. D. King\textsuperscript{1}

\textsuperscript{1}Department of Computer Science, University of Wales Aberystwyth, SY23 3DB, UK and \textsuperscript{2}Institute for Computer Science, Albert-Ludwigs-University Freiburg, D-79110 Freiburg, Germany

Received on August 19, 2005; revised on January 9, 2006; accepted on February 7, 2006
Advance Access publication February 15, 2006
Associate Editor: Nikolaus Rajewsky

ABSTRACT

Motivation: The genome of Arabidopsis thaliana, which has the best understood plant genome, still has approximately one-third of its genes with no functional annotation at all from either MIPS or TAIR. We have applied our Data Mining Prediction (DMP) method to the problem of predicting the functional classes of these protein sequences. This method is based on using a hybrid machine-learning/data-mining method to identify patterns in the bioinformatic data about sequences that are predictive of function. We use data about sequence, predicted secondary structure, predicted structural domain, InterPro patterns, sequence similarity profile and expressions data.

Results: We predicted the functional class of a high percentage of the Arabidopsis genes with currently unknown function. These predictions are interpretable and have good test accuracies. We describe in detail seven of the rules produced.

Availability: Rulesets are available at http://www.aber.ac.uk/compsci/Research/bio/dss/arabpreds/ and predictions are available at http://www.genepredictions.org

Contact: afc@aber.ac.uk

1 INTRODUCTION

Arabidopsis thaliana is the most important ‘model system’ in plant biology. Its genome was the first of a plant to be sequenced in December 2000 (The Arabidopsis Genome Initiative, 2000). Arabidopsis was chosen because of its amenable properties for experiments: rapid growth, small size and prolific offspring production. Understanding its genome is a basis for understanding the biology of all plants, and the crops we depend on for life.

At the time the sequence was published, 69% of genes were assigned to functional categories by sequence similarity. Most of these functional assignments were produced automatically using sequence similarity to other proteins of ‘known’ function, with only 9% being characterized experimentally. Thus, ~30% remained unclassified, have no close similar sequences of known function.

Although there are now more annotations that have been examined manually since the original annotation, the proportion of genes still with no annotation at all has not changed. The Arabidopsis Information Resource (TAIR: http://www.arabidopsis.org/) provides annotations to Gene Ontology (Gene Ontology: http://www.geneontology.org/) terms. As of March 2, 2004, TAIR had no annotation or annotation to ‘molecular function unknown’ for 35% of genes. This figure changes little over time.

Using the TAIR/TIGR annotations from GeneOntology.org as of January 2, 2006, 37% of genes had no annotation, or annotation to ‘molecular function unknown’. As of March 3, 2004, MIPS (MIPS: http://mips.gsf.de/) had no annotation or annotation to ‘unclassified proteins’ or ‘classification not yet clear-cut’ for 44% of genes. By January 2, 2006, this figure for the MIPS Pedant automatically derived categories was 58%. Gutierrez et al. (2004) estimate that 3848 Arabidopsis proteins are plant specific, and more than half of these are of unknown function.

The use of computational methods in making predictions for gene function is now a well-established field (Attwood, 2000; Hvidsten 2001; Marcotte et al., 1999; Pavlidis et al., 2001; Syed and Yona, 2003). We have previously successfully used machine learning and data mining to predict functions of genes in Saccharomyces cerevisiae, Escherichia coli and Mycobacterium tuberculosis (Clare and King, 2003b; King et al., 2000, 2001), and many of these predictions have been confirmed since they were made (King et al., 2004).

In this paper we use machine learning and data mining to produce rules that predict the function of genes in Arabidopsis from a wide variety of sources of data. This is a challenging bioinformatic task because the assembled data are often relational and highly structured, and cannot readily be used in standard attribute-value data mining/machine learning approaches. We use a combination of multi-relational data mining and the conventional decision tree learning algorithm C4.5 (Quinlan, 1993).

We use several different types of bioinformatic data. The different data sources are used independently and then as a single combined data source, so that we can see their relative potential and combined ability. The functional classification schemes, data, methods and results are described in more detail in the following sections.

2 FUNCTION

Functional classification schemes are now widely used in genome annotation. Probably the first such scheme was developed specifically for E.coli (Riley, 1993), and since then a number of similar schemes have arisen. The two most widely used function classification schemes are Gene Ontology (GeneOntology Consortium, 2000, http://www.geneontology.org/) and MIPS’s FunCat (FUNCAT: http://mips.gsf.de/proj/funcatDB/search_main_frame.html, Frishman et al., 2001). Each of these schemes is essentially a tree ordered by generality of function, with general classes at the top of the hierarchy which are broken down into more specific classes further down the hierarchy (controversially Gene Ontology is not a pure tree and small parts of it have the form of a directed
acyclic graph. Both schemes provide a controlled vocabulary of
terms to describe function.
However, neither scheme is ideal for use in the automated
prediction of functional class (Kell and King, 2000). The relations-
ships between terms and between levels of the hierarchy, the
orthogonality of concepts they describe, the completeness and con-
sistency of terms and the deductions that can be made from the
relationships still all need to be clarified. However, such schemes
are an undeniable step forward and enable new approaches to
functional bioinformatics, such as those applied in this paper.

Annotations to classes in Gene Ontology have been made avail-
able for at least 20 different species, sometimes from more than one
annotating organization. Annotations to MIPS’s FunCat scheme
have been made for 10 species. The FunCat originated from the
need to annotate S. cerevisiae, but has recently been expanded to be
suitable for other species. MIPS maintains a mapping between GO
and MIPS classes.

We chose to distinguish between annotations that had been
made automatically and those that had been made manually.
Gene Ontology provides an evidence code for each annotation
indicating how the annotation was derived. All annotations with
the code ‘IEA’ (Inferred by Electronic Annotation) were taken to
be automatic, and all others manual. MIPS provides two independent
sets of annotations, which are described as ‘automatic’ and ‘manual’.

We chose to use the top four levels of Gene Ontology’s
‘molecular_function’ scheme and the top four levels of MIPS’s
FunCat.

3 DATA
Six sources of data were used: sequence, expression, SCOP,
secondary structure, InterPro and sequence similarity. In addition,
a composite dataset was formed.

3.1 Sequence
The sequence data comprised attributes that could be calculated
directly from the sequence alone or data relating to the position
of the sequence in the genome. These were either obtained from
MIPS (MIPS Arabidopsis FTP: ftp://ftp.mips.gsf.de/cress/) or
Expasy’s ProtParam tool (Expasy: http://www.expasy.ch/) or
calculated directly. The attributes are summarized in Table 1. There
were a total of 4450 propositional attributes for each protein
sequence.

3.2 Expression
The expression data were provided by NASC’s (NASC: http://nasc.
nott.ac.uk/) Affymetrix service ‘Affywatch’ (Affywatch: http://
affymetrix.arabidopsis.info/AffyWatch.html). We used the results
of 43 experiments from CDs provided between December 2002
and January 2004, taking the signal, detection call and detection
P-values. This gave a dataset consisting of 1251 attributes. The
Affywatch service uses Affymetrix Arabidopsis 8 K and 25 K
arrays, so some data are available for the majority of the protein
sequences in the genome, and some for only a subset.

3.3 Predicted SCOP class
The SCOP data comprise the SCOP superfamily class predictions as
made by the Superfamily server (Gough et al., 2001, Superfamily:
http://supfam.mrc-lmb.cam.ac.uk/SUPERFAMILY/). This server

<table>
<thead>
<tr>
<th>Table 1. Sequence data</th>
</tr>
</thead>
<tbody>
<tr>
<td>Attribute</td>
</tr>
<tr>
<td>Strand</td>
</tr>
<tr>
<td>Chromosome</td>
</tr>
<tr>
<td>startpos</td>
</tr>
<tr>
<td>endpos</td>
</tr>
<tr>
<td>numpos</td>
</tr>
<tr>
<td>numaa</td>
</tr>
<tr>
<td>mol_wt</td>
</tr>
<tr>
<td>theo_pl</td>
</tr>
<tr>
<td>percentneg</td>
</tr>
<tr>
<td>percentpos</td>
</tr>
<tr>
<td>C,H,N,O,S</td>
</tr>
<tr>
<td>instability</td>
</tr>
<tr>
<td>aliphatic</td>
</tr>
<tr>
<td>gravy</td>
</tr>
<tr>
<td>seq_len</td>
</tr>
<tr>
<td>amino acid ratios</td>
</tr>
</tbody>
</table>

Attributes in the top section of the table were obtained from MIPS and were accessed
on March 16, 2003. Attributes in the middle section came from Expasy’s ProtParam tool
and were calculated on May 15, 2003. Attributes in the lower section were calculated
directly.

3.4 Predicted secondary structure
Secondary structure was predicted for each protein sequence
using Prof (Ouali and King, 2000, Prof: http://www.aber.ac.uk/~phiwww/prof/). This gives a prediction of alpha helix, beta strand
or coil for each amino acid position in the protein sequence. These
are then represented as a set of Prolog facts describing the positions
and lengths of each alpha, beta or coil section. The Prolog facts were
then mined by the first order association mining algorithm Poly-
FARM (Clare and King, 2003a, a Warmr-like algorithm, in order to
collect frequently occurring associations of facts. A total of 14806
frequently occurring associations were collected from the total of
1292068 training data facts, using a minimum frequency of 0.01.
The frequency threshold was experimentally chosen to provide a
tractable number of attributes. Boolean attributes were then formed
depending on whether these patterns occurred in the sequence
or not. These attributes are the final dataset. For this type of data
constructed in this manner, a typical C4.5 training data file is
183 Mb in size.
Sequence "Atlg01150" crc64 checksum: 95CFF5F90B6FB300 length: 315 aa.

<table>
<thead>
<tr>
<th>InterPro method</th>
<th>AccNumber</th>
<th>shortName</th>
<th>location</th>
</tr>
</thead>
<tbody>
<tr>
<td>IPRO001005</td>
<td></td>
<td>Myb DNA-binding domain</td>
<td></td>
</tr>
<tr>
<td>ProfileScan</td>
<td>PS00090</td>
<td>MYB_3</td>
<td>T[207–267]</td>
</tr>
<tr>
<td>ScanRegExp</td>
<td>PS00037</td>
<td>MYB_1</td>
<td></td>
</tr>
<tr>
<td>InterPromethod</td>
<td>IPRO001965</td>
<td>Zn-finger-like, PHD finger</td>
<td></td>
</tr>
<tr>
<td>ScanRegExp</td>
<td>PS01359</td>
<td>ZF_PHD_1</td>
<td>? [65–107] NA</td>
</tr>
<tr>
<td>InterPromethod</td>
<td>NULL</td>
<td>NULL</td>
<td></td>
</tr>
<tr>
<td>Coi1</td>
<td>AccNumber</td>
<td>shortName</td>
<td>location</td>
</tr>
<tr>
<td>Seg</td>
<td>seg</td>
<td>coiled-coil</td>
<td>? [168–189]</td>
</tr>
</tbody>
</table>

Fig. 1. Example of a sequence with three InterPro records, which can each consist of signatures from several databases.

### 3.5 InterPro

InterPro is a collection of several motif or signature finding databases. These include PROSITE, PRINTS, Pfam, ProDom, SMART and TIGRFAMs databases. InterPro data were calculated using the EBI’s stand-alone InterProScan package (Zdobnov and Apweiler, 2001, InterProScan: ftp://ftp.ebi.ac.uk/pub/databases/interpro/iprscan/), and the tool and collection of databases were downloaded on June 17, 2003. Signatures from several databases can be grouped within one InterPro entry, if they refer to the same protein family. Thus the data are easily described in a relational manner (Džeroski and Lavrač, 2001; ACM, 2003)—a protein sequence is found to have zero or more InterPro entries, each of which has one or more component database entries (i.e. hits in the protein motif databases). An example of a sequence with three InterPro entries is given in Figure 1. The InterPro data were then mined in the same manner as the structure data, in order to produce frequent patterns that could be used as Boolean attributes. A total of 82,722 training data facts became 2,817 Boolean attributes. A typical C4.5 training file is 36 Mb.

### 3.6 Sequence similarity profile

A PSI-BLAST search (Altschul et al., 1997) was conducted for each protein sequence against NCBI’s NRDB (non-redundant protein database). We used blastp (BLASTP 2.0.6), and nrdb as of May 24, 2001, ftp://ftp.ncbi.nlm.nih.gov/blast/db/ (maximum 20 iterations, –h0.0005, –e500). For each Arabidopsis sequence we wished to make use of as much of the information known about their similar sequences as possible. Therefore, all hits with E-values less than 10 that have a SWISS-PROT accession id were retrieved from SWISS-PROT v4.1. The SWISS-PROT information about these hits was translated into Prolog facts. The types of facts used for each hit are shown in Table 2.

The sequence similarity profile data were extremely large, consisting of more than 117 million facts in the training data alone. This was mined using PolyFARM, in the same way that the structure data were mined. A total of 197,539 frequent associations were produced (mining with a minimum frequency of 0.005). If each Arabidopsis sequence has nearly 200,000 Boolean attributes, this can create very large files. Even when the data are split into training, validation, test and ‘unknown’ sets we can be dealing with files that approach 2 Gb in size, when each Boolean is stored as a character, (as is normal for input into machine learning algorithms). In order to reduce the number of attributes, they were filtered to exclude attributes that remained constant across 95% of the training data. This gave us a set of 72,871 attributes. A typical C4.5 training data file is still 819 Mb.

### 3.7 Composite

In addition to the individual datasets, a composite dataset was formed by combining the six datasets. Feature selection was applied to each to reduce the volume of data. This was done by filtering features that remained constant across some percentage of the training data. The percentage used varied across the datasets, depending on their size, and ranged from 80% for the sequence similarity data up to 98% for the InterPro and SCOP data. The composite dataset contained 42,704 attributes.

### 4 METHODS

We used the Data Mining Prediction (DMP) method as described in more detail in King et al. (2000). Briefly the data are split into two-third for model-building and one-third to be held out as independent test data. The model-building data are again further split into two-third for training and one-third for validation. If the data are relational a preprocessing step of first association mining [PolyFARM (Clare and King, 2003a)] was applied to the model-building data. The frequent associations found in the data are used as Boolean attributes (testing whether each sequence has this association or not).

After the data had been propositionalized, C4.5 (Quinlan, 1993), a decision tree algorithm, was applied to the training data to produce rules. We applied our modified version of C4.5 that allows sequences to be labeled on their size, and ranged from 80% for the sequence similarity data up to 98% for the InterPro and SCOP data. The composite dataset contained 42,704 attributes.

The rules were tested against the validation data in order to filter out rules that overfitted the data. We tested for statistical significance of the precision or not).

After the data had been propositionalized, C4.5 (Quinlan, 1993), a decision tree algorithm, was applied to the training data to produce rules. We applied our modified version of C4.5 that allows sequences to be labeled with multiple functional classes (Clare and King, 2002).

The rules were tested against the validation data in order to filter out rules that overfitted the data. We tested for statistical significance of the precision or not).

Finally we measured the precision (positive predictive value) of the selected rules against the held-out test data, and applied the rules to the protein sequences of unknown function in order to make predictions of their function.

This process was repeated for each type of annotation (MIPS or GO, manual or automatic), and for each type of data, for comparison purposes.
The modified version of C4.5 allows all functional categories to be dealt with at the same time.

5 RESULTS

The final rules produced have both good test precision and are biologically informative.

The full rulesets can be found at http://www.aber.ac.uk/compsci/Research/bio/dss/arabpreds/ and the predictions that have been made can be found in an easy-to-query and read format at http://www.genepredictions.org/

5.1 Precision and scale of predictions

In Tables 3–6 we detail our results. We tabulate the numbers of rules produced and their average test precision for the different annotation schemes, the different levels of the functional hierarchies and the different data types. We also detail the number of predictions we have made for the protein sequences of unknown function.

We report precision (positive predictive value, TP/TP + FP). This is because the measurement of true or false negatives is misleading in this type of machine learning, for two reasons. One is that our validation step is designed to selects for only the best rules, rather than complete coverage, so measuring negative predictions does not provide any insight into performance. The second is that the true/false negatives would far outweigh the few positives and detract from understanding of the results.

Table 3 shows the average rule precision and number of rules produced for each data type. The precision ranges between 42 and 85%. Expression data were the least informative, and predicted SCOP class data gave the most accurate rules.

The expression data were made up of data from many different experiments (see Section 3). Each experiment had been conducted for a particular purpose, such as analysis of the pho3 mutation, changes induced by cold in freezing sensitive mutants, response to non-metabolized glucose analogues, etc. and each experiment has just a few readings for each gene in the genome, usually between 6 and 10. The problem in using this data for functional bioinformatics is that the goals of such targeted experiments may not match particularly well with the functional classes that we wished to predict. The result is that very few rules are produced from the expression data. Hence, a single rule with poor precision can really lower the average precision of the ruleset produced. However, as noticed before in our studies on the bacterial genomes and yeast, expression data are a strong predictor of ribosomal proteins, achieving rules that have a precision of 96% on test data. Many more expression datasets are now available for Arabidopsis (more than 800 experiments on NASCArray as of January 2005), and this increased amount should improve the results when using this type of data in future.

Table 2. Facts used in sequence similarity data

<table>
<thead>
<tr>
<th>Fact</th>
<th>Description</th>
</tr>
</thead>
<tbody>
<tr>
<td>eval(Seq,SPId,EVal)</td>
<td>Seq is similar to a Swiss-Prot protein with accession number SPId, and E-value EVal</td>
</tr>
<tr>
<td>sq_len(SPId,Length)</td>
<td>Swiss-Prot protein SPId has sequence length Length</td>
</tr>
<tr>
<td>mol_wt(SPId,Weight)</td>
<td>Swiss-Prot protein SPId has molecular weight Weight</td>
</tr>
<tr>
<td>species(SPId,Species)</td>
<td>Swiss-Prot protein SPId belongs to Species</td>
</tr>
<tr>
<td>classification(SPId,Classfn)</td>
<td>Swiss-Prot protein SPId belongs to Classfn genus</td>
</tr>
<tr>
<td>desc(SPId,Desc)</td>
<td>Swiss-Prot protein SPId has description Desc</td>
</tr>
<tr>
<td>keyword(SPId,Keyword)</td>
<td>Swiss-Prot protein SPId has keyword Keyword</td>
</tr>
<tr>
<td>db_ref(SPId,Database)</td>
<td>Swiss-Prot protein SPId has a reference in Database</td>
</tr>
</tbody>
</table>

The modified version of C4.5 allows all functional categories to be dealt with at the same time.

Table 3. Total number of rules produced and average rule precision (across all 4 hierarchy levels) 'GO all' are the GO annotations made by TAIR including those with IEA evidence codes, 'GO manual' are those excluding IEA annotations. 'MIPS auto' are MIPS FunCat automatic annotations, 'MIPS manual' are manual annotations.

<table>
<thead>
<tr>
<th>Data type</th>
<th>GO all</th>
<th>GO manual</th>
<th>MIPS auto</th>
<th>MIPS manual</th>
</tr>
</thead>
<tbody>
<tr>
<td>Number of rules produced</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>seq</td>
<td>105</td>
<td>104</td>
<td>55</td>
<td>19</td>
</tr>
<tr>
<td>expr</td>
<td>9</td>
<td>9</td>
<td>5</td>
<td>6</td>
</tr>
<tr>
<td>scop</td>
<td>362</td>
<td>301</td>
<td>478</td>
<td>49</td>
</tr>
<tr>
<td>struc</td>
<td>164</td>
<td>182</td>
<td>112</td>
<td>30</td>
</tr>
<tr>
<td>interpro</td>
<td>227</td>
<td>181</td>
<td>223</td>
<td>37</td>
</tr>
<tr>
<td>sim</td>
<td>331</td>
<td>295</td>
<td>430</td>
<td>48</td>
</tr>
<tr>
<td>composite</td>
<td>221</td>
<td>183</td>
<td>237</td>
<td>37</td>
</tr>
<tr>
<td>Average rule precision (%)</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>seq</td>
<td>56</td>
<td>57</td>
<td>52</td>
<td>55</td>
</tr>
<tr>
<td>expr</td>
<td>42</td>
<td>45</td>
<td>49</td>
<td>56</td>
</tr>
<tr>
<td>scop</td>
<td>82</td>
<td>85</td>
<td>72</td>
<td>79</td>
</tr>
<tr>
<td>struc</td>
<td>63</td>
<td>67</td>
<td>55</td>
<td>66</td>
</tr>
<tr>
<td>interpro</td>
<td>69</td>
<td>71</td>
<td>54</td>
<td>68</td>
</tr>
<tr>
<td>sim</td>
<td>77</td>
<td>79</td>
<td>66</td>
<td>66</td>
</tr>
<tr>
<td>composite</td>
<td>77</td>
<td>79</td>
<td>64</td>
<td>71</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Level</th>
<th>GO all</th>
<th>GO manual</th>
<th>MIPS auto</th>
<th>MIPS manual</th>
</tr>
</thead>
<tbody>
<tr>
<td>Number of rules produced</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>1</td>
<td>248</td>
<td>212</td>
<td>239</td>
<td>75</td>
</tr>
<tr>
<td>2</td>
<td>412</td>
<td>363</td>
<td>515</td>
<td>86</td>
</tr>
<tr>
<td>3</td>
<td>382</td>
<td>363</td>
<td>494</td>
<td>40</td>
</tr>
<tr>
<td>4</td>
<td>377</td>
<td>317</td>
<td>292</td>
<td>25</td>
</tr>
<tr>
<td>Average rule precision (%)</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>1</td>
<td>80</td>
<td>81</td>
<td>69</td>
<td>76</td>
</tr>
<tr>
<td>2</td>
<td>74</td>
<td>76</td>
<td>61</td>
<td>64</td>
</tr>
<tr>
<td>3</td>
<td>72</td>
<td>73</td>
<td>66</td>
<td>64</td>
</tr>
<tr>
<td>4</td>
<td>71</td>
<td>74</td>
<td>66</td>
<td>71</td>
</tr>
</tbody>
</table>

‘Level’ represents the level in the classification hierarchy: 1 is most general, 4 is most specific. ‘GO all’ are the GO annotations made by TAIR including those with IEA evidence codes, ‘GO manual’ are those excluding IEA annotations. ‘MIPS auto’ are MIPS FunCat automatic annotations, ‘MIPS manual’ are manual annotations.
Table 5. Number of predictions made (by level or by datatype)

<table>
<thead>
<tr>
<th>Level</th>
<th>GO all</th>
<th>GO manual</th>
<th>MIPS auto</th>
<th>MIPS manual</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>7587</td>
<td>12298</td>
<td>1737</td>
<td>7637</td>
</tr>
<tr>
<td>2</td>
<td>3355</td>
<td>6002</td>
<td>1074</td>
<td>4793</td>
</tr>
<tr>
<td>3</td>
<td>2821</td>
<td>6134</td>
<td>1755</td>
<td>3296</td>
</tr>
<tr>
<td>4</td>
<td>4678</td>
<td>9042</td>
<td>2557</td>
<td>14894</td>
</tr>
</tbody>
</table>

Data type
- seq: 16000
- expr: 5
- scop: 130
- struc: 130
- interpro: 330
- sim: 130
- composite: 130

Table 6. Total number of predictions made with confidence of 75% or greater using training data from each classification scheme for genes whose function is unknown to each classification scheme and for genes whose function is unknown to any scheme

<table>
<thead>
<tr>
<th>GO all</th>
<th>GO manual</th>
<th>MIPS auto</th>
<th>MIPS manual</th>
<th>Combined</th>
</tr>
</thead>
<tbody>
<tr>
<td>Unknown to this scheme</td>
<td>10532</td>
<td>15586</td>
<td>4574</td>
<td>18500</td>
</tr>
<tr>
<td>Unknown in all schemes</td>
<td>5572</td>
<td>7769</td>
<td>1364</td>
<td>5977</td>
</tr>
</tbody>
</table>

Total also shown for the combination of all rules produced from using all schemes.

The success of the SCOP data indicates the importance of structure in determining protein function (Thornton, 2001). These data produced many rules, and these rules usually contain only a single precondition and are therefore testing for a match to a single SCOP class. This indicates a very close agreement between the functional class hierarchies and the SCOP hierarchy. It also neatly illustrates C4.5’s ability to collapse complicated trees into simple rules.

The composite dataset performs slightly worse than the SCOP data and approximately the same as the sequence similarity data. This indicates an inefficiency in the use of the composite information. The rules show a good mix of preconditions from all the data types (with expression data preconditions underrepresented, again showing that these are not strong class discriminators).

The figures for the different levels of the hierarchy in Table 4 indicate how the predictions change as the functional classes become more specific. Precision is highest at the most general level, but does not vary much across the levels.

Many predictions have been made, and the numbers of these predictions can be seen in Tables 5 and 6. Table 5 shows the total number of predictions broken down by datatype and by level. We can see that the expression data produce rules that make the fewest predictions (in line with it producing the fewest rules), and the sequence similarity and composite data make the most predictions. This perhaps indicates that rules produced from the composite data capture information about classes whose genes are as yet not well determined. Table 6 shows the total number of predictions that were made from these rules for genes of unknown function. We made predictions with an estimated precision of at least 75% for a total of 8156 protein sequences with no annotation from either GO/TAIR or MIPS.

5.2 Example analysis of several rules

In this section we describe a few of the rules in order to demonstrate their typical structure and interpretation. They are shown to be consistent with known biology and are suggestive of novel science.

The rule in Figure 2 states that if the lysine–arginine ratio where these 2 amino acids are 7 residues apart (i.e. there are 6 residues between them) is greater than 1.485, then the gene is involved in protein synthesis. This rule has a precision of 42% on the test data (8/19), whereas the prior probability of this class was just 4%. Of the 11 errors of commission, 5 are histone proteins. In our previous work on the M. tuberculosis genome (King et al., 2000) we found a similar rule covering genes involved in ribosomal protein synthesis, with the single precondition that the protein has >6.6% lysine. Both lysine and arginine are positively charged and are together known to make up to 29% of a histone protein, concentrated in the ‘arms’ of the proteins that bind to the negatively charged phosphate groups in DNA. Also, lysine and arginine are abundant in the nuclear localization signal. This is a particular motif that allows certain proteins to be transported through the nuclear envelope (see the NLSdb, A database of Nuclear Localization Signals: http://cubic.bioc.columbia.edu/db/NLSdb/, for examples of this motif). Such proteins include histones and ribosomal proteins (moving into the nucleus after being produced in the cytosol). Interestingly, the training data proteins that match this rule do not align well using ClustalX. There are several strong pairwise alignments, but otherwise the collection is diverse. The K.....R segments are scattered across the length of the proteins. The six residues between the lysine and arginine varied and an alignment of all the isolated K.....R motifs showed no trends, but the most frequently occurring were aliphatics (L, A and V) and more lysine and arginine.
then dbref the function is GO:0004803 'transposase activity' not highly similar to a short hydrolase protein and can activate its own transposition. The Mutator autonomous et al. (Eisen Mutator is the most active transposable system in higher plants required for transposition of the Mutator family of transposons. Blast Web Service detected putative conserved domains for Mutator 14 matching test set sequences to be very strongly related. NCBI's involved in transposase activity. A ClustalX alignment shows the Arabidopsis have none. The role of most of the serine carboxypeptidases in carboxypeptidases, but not highly similar to a lactobacillus protein but not highly similar to a bacillus subtilis protein and not highly similar to a short hydrolase protein then the function is GO:0004803 'transposase activity'

labeled as ‘structural constituent of ribosome’. The errors tend to be DNA binding proteins. This time the rule requires a high arginine–valine ratio, and a high theoretical pl (isoelectric point). The isoelectric point is the pH at which the protein would have no net charge. Having a theoretical pl above 10 means that at a neutral pH the protein will be positively charged (whereas most proteins would be negatively charged at a neutral pH). This is consistent with the protein binding to DNA/RNA and with the rule in Figure 2.

The rule in Figure 4 has a precision of 100% on all datasets (training, test and validation), and the matching sequences (38, 18 and 35, respectively) have either serine carboxypeptidase activity or ubiquitin conjugating enzyme activity. The Pichia are yeast, a genus of fungi, similar to Candida. ‘hssp dbref’ refers to the protein having an entry in the HSSP database (HSSP: http://www.cmbi.kun.nl/gv/hssp/) (Homology-derived Secondary Structure of Proteins). The Proteobacteria are a diverse class of Gram-negative bacteria containing nitrogen-fixing bacteria and enterobacteria such as E.coli. This rule makes predictions for seven genes of unknown function. Ubiquitin is found in all eukaryotes, but not in bacteria, which is consistent with this rule. The Merops (Merops: http://merops.sanger.ac.uk/) peptidase database summary for peptidase family S10 shows that while Arabidopsis is known to have none. The role of most of the serine carboxypeptidases in Arabidopsis is unknown (Lehfeldt et al., 2000).

The rule in Figure 5 has a precision of 100% on the test data and makes predictions for 37 sequences of unknown function to be involved in transposase activity. A ClustalX alignment shows the 14 matching test set sequences to be very strongly related. NCBI’s Blast Web Service detected putative conserved domains for Mutator transposases in these sequences. Mutator transposases are factors required for transposition of the Mutator family of transposons. Mutator is the most active transposable system in higher plants (Eisen et al., 1994) and the Mutator system is autonomous—it can activate its own transposition. The Mutator autonomous element contains two protein sequences encoding the proteins MURA and MURB. MURA is believed to be the transposase and has similarity to several bacterial insertion sequences, including IS256 (Eisen et al., 1994). IS256 is found in Lactobacillus but not in Bacillus (Mahillon and Chandler, 1998) (Table 3 at http://mmbr.asm.org/cgi/content/full/62/3/725/T3). This rule raises interesting questions about the origin of plant transposases.

Figure 6 shows a simple rule, which has a precision of 100% on all three datasets (training, validation and test) and makes predictions for 4 genes of unknown function. Why these particular conditions should hold is unclear. The Lymnaeidea are a superfamily of pond snails, including Biomphalaria glabrata (the most important intermediate hosts for a widespread pathogen of humans, the digenetic trematode Schistosoma mansoni). The genus Rattus encompasses the rat species. Smart is a database that contains signatures from domains from signaling and extracellular proteins. Proteins that match this rule are all very similar to a protein from Biomphalaria glabrata [SWISS-PROT p92179]. This is a cytoplasmic actin. These are only weakly similar to Rattus proteins, so it would appear that vertebrate lineage has diverged more strongly from the ancestral form of the protein than either plants or molluscs. This implies evolutionary pressure on these proteins in mammals. The genes of unknown function predicted by this rule are At1g03060 (WD-40 repeat family protein/beige-related), At1g61180 (disease resistance protein), At4g16960 (disease resistance protein) and At5g43500 (which encodes 3 potential proteins including one whose ‘sequence is similar to actin-related proteins (ARP’s) in other organisms’).

The rule in Figure 7 has a precision of 97.14% (68/70) on the held out test set, and is based solely on a single InterPro record. This makes predictions for 45 genes which are either of previously unknown function, or annotated less specifically, such as ‘nucleic acid binding’. This highlights an inefficiency in previous annotations.
The rule shown in Figure 8 has a precision of 71% on test data and makes many predictions for genes of unknown function. SCOP class ‘e’ is the class of multi-domain proteins, proteins with folds consisting of two or more domains belonging to different classes. *Viridiplantae* is the class of green plants. Here the multi-domain condition is intriguing.

### 6 DISCUSSION AND CONCLUSIONS

We have demonstrated the value of the DMP method of predicting protein functional class across the full spectrum of genome types, from bacterial up to the genomes of multi-cellular eukaryotes such as *Arabidopsis* (King et al., 2000; Clare and King, 2003a).

A number of general conclusions can be made from these studies:

- It is possible to learn accurate rules that are more general that can be produced using traditional sequence similarities.
- The rules produced can, at least to a limited extent, be biologically interpreted. This demonstrates that they are consistent with known biology, and that they are capable of suggesting new ideas.
- Sometimes, the rules can be difficult to interpret; even when shown to be accurate. This highlights the distinction in the philosophy of science between ‘prediction’ and ‘explanation’.
- Some rules appear consistently across genomes, such as the rule that a high lysine content is an indicator of ribosomal protein biosynthesis.
- The different sources of information are of varying value. The sequence similarity profile data are perhaps the richest and most valuable source of data. Structural information is also of great use. Expression data appear to be the least informative. The probable reason for this is that expression experiments are conducted with a specific question in mind. Despite this, expression data are a strong discriminator of ribosomal proteins.

The data that we collected for our equivalent study of the yeast genome (Clare and King, 2003a) have now been used to provide the basis for the ILP Conference Challenge in 2005 (15th International Conference on Inductive Logic Programming http://ilp2005.in.tum.de/). This is a data mining challenge for multi-relational data. This means that not only are the gene predictions and rules a valuable resource for functional genomics, but the data itself are inspiring developments in the computational field of multi-relational data mining. We expect that the data, rules and predictions we have collected for *Arabidopsis* will prove equally valuable to others.

All the predictions produced are publicly available, together with links to the rules and their estimated precision, at http://www.genepredictions.org/. We are confident that many of these will be shown to be correct in the future, and we are currently experimentally testing insertion mutants of several of our predictions.

### ACKNOWLEDGEMENTS

A.C. and A.K. received support from the BBSRC grant B/H/14248. The work was also partially supported by the IO project (EU grant IST-FET FP6-516169).

### Conflict of Interest

none declared.

### REFERENCES


