BIOINFORMATICS

Vol. 17 no. 5 2001
Pages 445–454

The utility of different representations of protein sequence for predicting functional class

Ross D. King [1,], Andreas Karwath [1], Amanda Clare [1] and Luc Dehaspe [2]*

[1]Department of Computer Science, University of Wales, Aberystwyth, Penglais, Aberystwyth, Ceredigion SY23 3DB, Wales, UK and [2]PharmaDM, Ambachtenlaan 54, B3-3001 Leuven, Belgium

Received on October 17, 2000; revised and accepted on January 19, 2001

ABSTRACT

Motivation: Data Mining Prediction (DMP) is a novel approach to predicting protein functional class from sequence. DMP works even in the absence of a homologous protein of known function. We investigate the utility of different ways of representing protein sequence in DMP (residue frequencies, phylogeny, predicted structure) using the *Escherichia coli* genome as a model.

Results: Using the different representations DMP learnt prediction rules that were more accurate than default at every level of function using every type of representation. The most effective way to represent sequence was using phylogeny (75% accuracy and 13% coverage of unassigned ORFs at the most general level of function: 69% accuracy and 7% coverage at the most detailed). We tested different methods for combining predictions from the different types of representation. These improved both the accuracy and coverage of predictions, e.g. 40% of all unassigned ORFs could be predicted at an estimated accuracy of 60% and 5% of unassigned ORFs could be predicted at an estimated accuracy of 86%.

Availability: The rules and data are freely available. Warmr is free to academics.

Contact: rdk@aber.ac.uk

Supplementary information: http://www.aber.ac.uk/~dcswww/Research/bio/ProteinFunction

INTRODUCTION

The first step in determining the function of a newly sequenced protein is to attempt to predict its function using bioinformatic techniques. This is conventionally done using statistically based sequence similarity (SIM) methods which predict function based on inferred orthologous homology, e.g. FASTA (Pearson and Lipman, 1988) and PSI-BLAST (Altschul *et al.*, 1997). Such bioinformatic predictions make experimental determination of function simpler as it is clearly more efficient to test an accurate prediction than to randomly test for possible functions.

To predict protein function directly from sequence what is abstractly required is a discrimination function (Mitchell, 1997) which maps sequence to biological function. The existing sequence homology recognition methods can be considered as examples of nearest neighbour discrimination functions (Duda and Hart, 1973) in sequence space. Recognising that the problem of prediction function from sequence is a discrimination problem makes it clear that many other data analysis approaches can be applied to the problem.

We have recently developed a novel method of predicting function based on using data mining/machine learning to induce rules which map from sequence to functional class (King *et al.*, 2000a,b). We call this method Data Mining Prediction (DMP). The DMP approach has several advantages over conventional SIM methods:

- Function can be predicted in the absence of homology to a sequence of known function.
- More general types of SIM can be utilised allowing more remote homologies to be detected.
- Explicit comprehensible rules can be produced which may provide biological insight.

The disadvantages of DMP are:

- It requires standard SIM functional assignments to bootstrap from.
- It can only identify the functional class of a protein, not its specific function.

Suprisingly little work has been done on the prediction of function from sequence using methods other than direct SIM. The closest previous work to DMP was carried out by des Jardins *et al.* (1997), who used sequence based descriptors and machine learning to predict if a protein

*To whom correspondence should be addressed.

was an enzyme, and EC classification if it was known to be an enzyme. However, the authors did not demonstrate any advantages over conventional SIM based classification methods. Work on protein fold prediction (e.g. Jones *et al.*, 1992; Kelly *et al.*, 2000) is related to predicting function from sequence—especially where analogous folds are predicted. If a fold family is predicted for a sequence which has a member of known function, then this function could be inferred for the sequence (a nearest-neighbour approach). In the case of novel folds it still may be possible to infer function (Stawiski *et al.*, 2000).

Functional hierarchies

The recognition of the value of organising the functions in proteins into classes (Riley and Labedan, 1996) is one of the most important conceptual advances in functional genomics (Rison *et al.*, 2000). Functional hierarchies are essential for DMP as they enable the learning (inducing) of general rules which discriminate between different classes. Once learnt such rules can be used to predict the class of proteins of unknown functional class.

We selected for study the functional hierarchy of *Escherichia coli* from the Riley group http://genprotec.mbl.edu/start. *E.coli* has arguably the best characterised extant genome, and the Riley functional classification is in our opinion the most researched and thorough of all functional hierarchies. A further advantage is that the *E.coli* functional hierarchy has probably a higher percentage of functions known from direct experimentation of any organism. A typical example of the classification of a protein in this hierarchy is that of pyruvate formate lyase activating enzyme (B4379, yjjW). This has a level 1 (most general) class of 'Metabolism of small molecules', a level 2 class of 'Energy metabolism, carbon' and level 3 (most specific) class of 'Anaerobic respiration'.

Data mining prediction (DMP)

The basic approach of DMP is as follows (see also Figure 1):

(1) Retrieve the identified open reading frames (ORFs) (putative proteins) and their known functional assignments (note, that some ORFs will be shown not to code for proteins, and there are errors in annotation of function, Brenner, 1999)—both of which add 'noise' to the data mining process (Mitchell, 1997).

(2) Describe each ORF in the genome using a defined language—the descriptions are based solely on information which can be computed from the sequence.

(3) Use data mining (Piatetsky-Shapiro and Frawley, 1991; Fayyad *et al.*, 1996; Munakata, 1999) to identify frequent patterns in the descriptions of the ORFs. The Inductive Logic Programming (ILP) algorithm Warmr was used (Dehaspe *et al.*, 1998). The frequent patterns were then converted into binary attributes which describe the ORFs.

(4) Use machine learning (Mitchell, 1997) to learn rules which map from the attributes describing the sequences to their function.

(5) Use the learnt rules to (inductively) infer the functional classes of ORFs of unknown function.

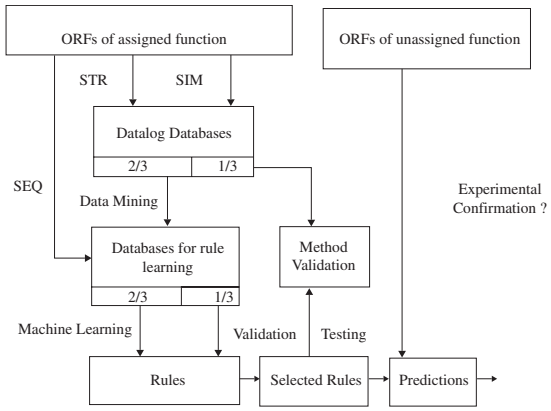

Fig. 1. Flow chart of the experimental methodology. Warmr is an Inductive Logic Programming (ILP) data mining algorithm. It used 2/3 of the data to learn frequent patterns. The remaining third was set aside as a final test set. The machine learning algorithm C5 was used to learn rules that predict function from the descriptional attributes using 4/9 of the data (2/3 of 2/3). Good rules were selected on the validation data—the remaining 2/9 of the dataset. The unbiased accuracy of these rules estimated on the 1/3 test set. The selection criteria for good rules was that on the validation data they covered at least two correct examples, had an accuracy of at least 50%, and an estimated deviation of $\geqslant 1.64$. This process was only carried out once because of computational resource limitations.

We have previously validated the DMP approach on both the *Mycobacterium tuberculosis*, and the *E.coli* genomes (King *et al.*, 2000a,b). On the *M.tuberculosis* genome the learnt rules predicted the functional class of 65% of the ORFs with no assigned function, and on the *E.coli* genome rules were learnt that predicted 24% of those with no assigned function. The rules had an estimated accuracy of 60–80% (depending on the level of functional assignment), and many of the rules were demonstrated to be not based on homology. The poorer result on *E.coli* than *M.tuberculosis* was due to the existing better knowledge of the *E.coli* genome and to it being less conservatively annotated.

In these previous experiments a wide variety of ways of describing protein sequences were used: residue frequency, phylogeny of identified homologues, predicted

Table 1. The sequence based attributes used to describe the ORFs. **R** is an amino-acid residue

Attributes	Description	No.	Type
amino_acids_R	The number of residues of type **R** in the sequence	21	int
amino_acid_ratio_R	The percentage composition of residues of type **R**	21	real
amino_acid_pairs_RS	The number of residue pairs of type **R**, **S** in the sequence	441	int
amino_acid_pair_ratio_RS	The percentage composition of residue pairs of type **R**, **S**	441	real
sequence_length	The number of residues in the sequence	1	int
molecular_weight	The computed molecular weight	1	int
aliphatic_index	The computed aliphatic index	1	real
hydro	The Grand average of hydropathicity (GRAVY), the value was discritised; 1 for low values, increasing up to 5 for high values	1	int
pI	The theoretical isoelectric point (pI) for this ORF	1	real
atomic_comp_E	The ORFs atomic composition of element **E**; where E is one of the following: carbon (C), hydrogen(H), nitrogen(N), oxygen(O), or sulfur(S)	5	int

There are considered to be 21 residues, the standard 20 plus x (for repetitive sequences, according to *pseg*; Wootton and Federhen (1993)). The last four attributes (aliphatic_index, hydro, pI, and atomic_comp_E) were generated using the ProtParam program (http://www.expasy.cbr.nrc.ca/tools/protparam.html).

secondary structure, etc.; and the rules generated were often a combination of these different descriptor types. It was therefore unclear what the relative importance of the different types of descriptor was. In this paper we focus on the question of the best way to describe protein sequences to infer function in DMP.

METHODOLOGY

Describing the ORFs

For *E.coli* we used the 4289 ORFs identified by Blattner *et al.* (1997) and took the functional assignments from the Riley group http://genprotec.mbl.edu/start. We considered the Riley classes 'Open Reading Frames' and 'Miscellaneous' to be 'unassigned'.

Three basic types of information were computed to describe protein sequences (ORFs):

- sequence based attributes (see Table 1);
- SIM (phylogeny) based Datalog descriptors (see Table 2),

Table 2. The phylogenic descriptors used

Database argument	Description
hom(P)	**P** is a homologous protein found by PSI-BLAST
e_val_rule(P, E)	**P** is a homologous protein found by PSI-BLAST with SIM measure **E**
e_val_lteq(P, X)	**P** is a homologous protein found by PSI-BLAST with SIM measure less than **X**
e_val_gt(P, X)	**P** is a homologous protein found by PSI-BLAST with SIM measure greater than **X**
psi_val_rule(P, It)	**P** is a homologous protein found by PSI-BLAST on iteration **It**
psi_iter_lteq(P, X)	**P** is a homologous protein found by PSI-BLAST on iteration less than **X**
psi_iter_gt(P, X)	**P** is a homologous protein found by PSI-BLAST on iteration greater than **X**
species(P, Species)	The protein **P** comes from species **Species**
classification(P, Class)	The protein **P** comes from a species with SwissProt phylogenic classification **Class**
mol_wt_rule(P, X)	The protein **P** has discretised molecular weight **X**
mol_wt_lteq(P, X)	The molecular weight of **P** is less than **X**
mol_wt_gt(P, X)	The molecular weight of **P** is greater than **X**
keyword(P, Word)	The SwissProt keyword **Word** describes protein **P**

These descriptors describe the result of PSI-BLAST sequence searches. The NRProt (05/10/99) database was used for maximum sensitivity, and the predicted homologous SwissProt (Bairoch and Apweiler, 1999) proteins extracted from it. The values described in the table by 'X' are discretised into 5 classes (1 very low, 2 low, 3 medium, 4 high, and 5 very high). The e value of a PSI-BLAST search is a measure of the probability of a sequence match being homologous (note that a low value means a high SIM). It can also be considered as a measure of evolutionary relatedness of the homologous protein. PSI-BLAST is an iterative search process which uses results from initial searches to guide later searches. The iteration in the search that a homologous protein is found is informative about the evolutionary relatedness of the homologous protein. To describe each homologous protein found we used the species name it was taken from and its complete phylogenic classification (Phylum–species). The keywords: membrane, transmembrane, inner_membrane, outer_membrane, repeat, plasmid, and alternative_splicing were also added to the database if they were present in the SwissProt description.

- predicted secondary structure based Datalog descriptors (see Table 3).

The sequence attributes (SEQ descriptors) are essentially based on the sequence's composition of singlets and pairs of residues (Table 1). The sequence-based attributes were directly calculated as attributes and were only used at the machine learning stage see Figure 1. Describing protein sequences using just the sequence composition of singlet and pairs of residues loses all information about the order of residues. However, previous results had suggested that this approach was surprisingly effective in deriving useful discriminatory sequence 'fingerprints' (King *et al.*, 2000a,b). There were 933 sequence based attributes.

Table 3. Database facts and their description

Database argument	Description
ss(S, T)	Position **S** is predicted to be a secondary structure element of type **T**
nss(S1, S2, T)	Given the secondary structure at position **S1**, the neighbouring position **S2**, with **S2** = **S1** + 1, has a secondary structure prediction of type **T**
ss_alpha(S, gt, X)	Position **S** is predicted to be an alpha-helix of length greater than **X** (similarly **lteq** instead of **gt**)
ss_beta(S, gt, X)	Position **S** is predicted to be a beta-strand of length greater than **X** (similarly **lteq** instead of **gt**)
ss_coil(S, gt, X)	Position **S** is predicted to be a coil of length greater than **X** (similarly **lteq** instead of **gt**)
nss_alpha(S1, S2, gt, B)	Positions **S1** and **S2** (where **S2** = **S1** + 2) are predicted to be alpha-helices, **S2** has length greater than **X** (similarly **lteq** instead of **gt**)
nss_beta(S1, S2, gt, X)	Positions **S1** and **S2** (where **S2** = **S1** + 2) are predicted to be beta-strands, **S2** has length greater than **X** (similarly **lteq** instead of **gt**)
nss_coil(S1, S2, gt, X)	Positions **S1** and **S2** (where **S2** = **S1** + 2) are predicted to be coils, **S2** has length greater than **X** (similarly **lteq** instead of **gt**)

These facts are generated for each of the genes. Positions in the text refer to the order in the predicted secondary structure. If for example an ORF has the following predicted secondary structure: $\alpha\alpha\alpha\alpha$ccccccc$\alpha\alpha\alpha\alpha\alpha$ccccccc$\beta\beta\beta$ would translate into: the 1st alpha-helix secondary structure prediction is of length 4; the 1st coil secondary structure prediction is of length 6; the 2nd alpha-helix secondary structure prediction is of length 5; the 2nd coil secondary structure prediction is of length 7; and the 1st beta-strand structure prediction is of length 3. The values described in the table by 'X' are discretised into 5 classes (1 very low, 2 low, 3 medium, 4 high, and 5 very high).

For each ORF in the *E.coli* genome we carried out a PSI-BLAST (Altschul *et al.*, 1997) SIM search (with parameters: $e = 10$, $h = 0.0005$, $j = 20$ with the NRProt 05/10/99 database). The result of these searches was used to calculate the SIM and predicted secondary structure based (STR) descriptors. (Note that PSI-BLAST can be confused by multidomain proteins and this will add noise to the data.) The SIM descriptors capture in the computer (logic) programming language Datalog (Ullman, 1988) the essential information in a PSI-BLAST result (Table 2). The use of a rich (first order) language such as Datalog allows the description to capture the distribution of the predicted homologous proteins to the ORF, their evolutionary distance from the target, the phylogenic relationship between the proteins, their relative sizes, and keywords describing the homologous sequences. The keywords are based on properties that can be predicted from sequence—especially the presence of membrane/trans-membrane binding sequences. Such a rich description would not be possible using conventional attribute vector descriptors. It is intuitively clear that SIM information can provide powerful clues to the function of a sequence even when the known function of homologous proteins is disregarded. For example, if a set of homologous proteins is found across the Eubacteria, Archaea, and Eukaryotes, then this increases the probability that they have an essential house-keeping role. The Datalog description of a PSI-BLAST run resembles the 'phylogenic profile' approach of the Eisenberg group (Marcotte *et al.*, 1999).

The ILP (Muggleton, 1991; Lavrac and Dzeroski, 1994) data mining method Warmr (Dehaspe *et al.*, 1998) was used to find frequent patterns in the Datalog descriptions of the PSI-BLAST runs. Warmr found 13 799 frequent patterns (see http://www.aber.ac.uk/~dcswww/Research/bio/ProteinFunction/ecoli.hom.out for the complete list). These frequent patterns were then converted into binary attributes (1 if the patterns was present in an ORF and 0 if the patterns was absent) for use by machine learning.

The descriptors based on the predicted secondary structure (STR) were also coded into Datalog (Table 3). The program Prof (Ouali and King, 2000) was used to make secondary structure predictions of all the *E.coli* ORFs. Prof uses as input the homologous sequences detected by PSI-BLAST, and employs cascaded neural networks to make the predictions. Prof has an estimated accuracy of 77%. Warmr found 18 342 frequent patterns in the secondary structure prediction data (see http://www.aber.ac.uk/~dcswww/Research/bio/ProteinFunction/ecoli.struc.out for the complete list). These, like the SIM frequent patterns, were converted into binary attributes (1 if the patterns was present in an ORF and 0 if the pattern was absent) for use in machine learning.

Basic learning experiments

The basic methodology of the learning experiments is shown in Figure 1. The experiments were designed to find rules that: given an ORFs sequence description, accurately predict its function. To minimise the danger of 'overfitting' rules to the data (finding rules that make accurate predictions on the data used to learn the rules, but low accuracy on new data), the ORFs of assigned function were split into three sets (training, validation, and test). Rules were first learnt on the training data and their performance examined on the validation set. Rules which performed well on the validation set were then selected, and their accuracy and coverage estimated on the test set (Mitchell, 1997). This method allows an unbiased estimate of accuracy and coverage.

This learning procedure was carried out for each of the three types of information about the ORFs (sequence based attributes—SEQ, SIM based Datalog

Table 4. The test set results for using the different types of sequence description (sequence based attributes—SEQ; sequence similarity based Datalog descriptors—SIM; predicted secondary structure based Datalog descriptors—STR; combined all rule sets—WTD_VOTE_ALL; combined all rules sets with prediction from $\geqslant$ 2 rules—VOTE_2_ALL; combined SEQ, SIM, STR only—WTD_VOTE_SSS; combined SEQ, SIM, STR with prediction from $\geqslant$ 2 rules—VOTE_2_SSS)

Attributes	Accuracy %			Coverage %			No. of Predictions		
	1	2	3	1	2	3	1	2	3
SEQ	64	63	41	20	18	4	359 (17)	245 (11)	63 (3)
SIM	75	74	69	29	26	16	290 (13)	288 (13)	152 (7)
STR	59	44	17	10	1	5	149 (7)	38 (2)	74 (3)
SEQ + SIM	84	71	60	23	28	16	231 (11)	272 (13)	115 (5)
SEQ + STR	69	64	50	20	22	3	317 (15)	401 (19)	37 (2)
SIM + STR	75	69	54	25	27	20	195 (9)	301 (14)	152 (7)
SEQ + SIM + STR	75	69	61	28	26	15	353 (16)	267 (12)	135 (6)
WTD_VOTE_ALL	60	54	42	52	48	36	863 (40)	818 (38)	475 (22)
VOTE_2_ALL	75	68	68	32	34	17	400 (18)	377 (17)	122 (6)
WTD_VOTE_ SSS	64	66	52	41	34	22	626 (29)	462 (21)	296 (14)
VOTE_2_ SSS	86	88	90	12	11	1	117 (5)	91 (4)	16 (1)

The numbers 1, 2, 3 correspond to the levels in *E.coli* functional hierarchy (level 1 the most general, level 3 the most specific). The test set accuracies are: (the number of ORF predicted to have the correct function/number of predictions) * 100. The default accuracies for the data (i.e. just choose the largest class) are for level one 40%, for level two 21%, and for level three 6%. The test coverages are: (the number of ORFs predicted to have a function/total number of ORFs) * 100. There were 712 ORFs of known function in the test set. The number and percentage (in brackets) of ORFs of unknown function (total of 2167) predicted by the rules from the different types of sequence description are given in the last three columns.

descriptors—SIM, predicted secondary structure based Datalog descriptors—STR). For each type of information we attempted to learn rules for every functional class in the three hierarchical levels of assigned function in *E.coli*.

Combined learning experiments

We also examined a number of different ways of combining the different type of information. The simplest of these was to combine the different types of attributes in the machine learning stage of the work. This produced four combinations: SEQ + SIM + STR, SEQ + SIM, SEQ + STR, SIM + STR. For each combination we attempted to learn rules for every functional class in the three hierarchical levels of assigned function.

In addition we investigated other ways of combining descriptional information based on ensemble learning (Bauer and Kohavi, 1999)—such approaches are state-of-the-art in machine learning. We found that the best results (the results we report here) came from a simple voting strategy—allowing each ruleset to vote for class of an ORF. Two different voting strategies are reported here VOTE_2 = only selecting predictions that are made by at least two rulesets, and WTD_VOTE = weighted voting, where the accuracy on the validation set is used to weight the vote of a rule. We first did this for the rulesets from all the types of description (SEQ, SIM, STR, SEQ+SIM, SEQ+STR, SIM+STR, SEQ+SIM+STR) which we term VOTE_2_ALL and WTD_VOTE_ALL. We also used both voting strategies with the just the basic descriptions (SEQ, SIM, STR) which we term VOTE_2_SSS and WTD_VOTE_SSS.

RESULTS

The test set accuracies and coverages of the rules are given in Table 4 (complete details can be found at http://www.aber.ac.uk/~dcswww/Research/bio/ProteinFunction).

It was possible to find rules that were more accurate than default at every level of function using every type of sequence description. Of the three basic types of description, SIM was the most effective. It gave both the highest accuracy and coverage at each level of function. This result agrees with intuition, as SIM is the richest of the three sequence description methods; there is also evidence that function can be predicted using 'phylogenic profiles' (Marcotte *et al.*, 1999) and these are related to the information in SIM. The sequence description methods SEQ and STR also perform creditably, with SEQ outperforming STR. Although SEQ used quite a simplistic way of representing sequences, based on the composition of singlets and pairs of residues, it is surprising how useful these compositional fingerprints were at predicting function. The power of SEQ for predicting function is one of the main results of this paper. STR also performed well: at the top level of function, STR predicted 10% of the test set with an estimated accuracy of 59%. We did not expect that it would be possible to predict functional class based solely on predicted secondary structure.

Using SIM it was possible to predict 15–30% of the test set with ~70% accuracy. We consider this to be remarkable. SIM includes only information gleaned only from the patterns of proteins found in SIM searches (of course excluding any explicit functional information in

such searches). SIM outperforms any other combination of descriptors (SEQ, SIM, STR, SEQ + SIM, SEQ + STR, SIM + STR, SEQ + SIM + STR), with the possible exception of SEQ + SIM. This result was unexpected, as SEQ and STR performed well on their own, and we anticipated that adding this information to the learning process should increase performance. It is possible that SEQ and SIM really add no information not present in SIM, but it is more likely that the result is due to inefficiencies in the machine learning process (e.g. too many attributes for C5 to deal with efficiently). The results for SEQ + STR (i.e. SIM excluded) are good at levels 1 and 2 (∼65% accuracy with ∼20% coverage), at level 3 only a poor coverage (3%) is obtained. It seems that the SEQ and STR descriptors are not rich enough to make fine functional discriminations.

The ensemble learning approaches of WTD_VOTE_ALL and WTD_VOTE_SSS had higher coverages than any other prediction methods. WTD_VOTE_ALL predicts the functional class of 40% of the ORFs at level 1 with an estimated accuracy of 60 and 22% of the ORFs at level 3 with and estimated accuracy of 42% (note default accuracy here is 6%). VOTE_2_ALL and VOTE_2_SSS, which only make predictions if rules agree from more than one descriptor type, had higher accuracies than those of WTD_VOTE_ALL and WTD_VOTE_SSS, again as expected. The predictions from VOTE_2_SSS were more accurate than those from VOTE_2_ALL as the descriptor types (SEQ, SIM, STR) are more independent than (SEQ, SIM, STR, SEQ + SIM, SEQ + STR, SIM + STR, SEQ + SIM + STR). VOTE_2_SSS is the most accurate prediction method developed with ∼90% accuracy at all levels.

As in previous work (King *et al.*, 2000a,b) we found that the predictions at the higher levels were more accurate and had higher coverage than low level predictions. This is to some extent expected, as there are many more functional classes to discriminate between at the lower levels (i.e. the prior probabilities of these classes are smaller). However, before we started work on DMP we did not believe that it was likely that it would be possible to discriminate between such broad classes as 'Cell processes' and 'Macromolecule metabolism', by inspection of their sequences. We believe this result is important biologically. The most valuable predictions are those at the lower levels as these can be tested most easily experimentally. At this level of detailed prediction it is relatively easy to envisage experiments to confirm the predictions. For example, the ORF B4082 (yjcR) is predicted to be in the functional class 'Chemotaxis and mobility'. Therefore, if this gene was knocked out and the cells displayed defects in chemotaxis and mobility, the experimental result would be consistent with the hypothesis.

Table 5. The number of selected rules from the different types of sequence description, and the percentage of non-homology based rules from the different types of sequence description

Attributes	No. of rules			% Non-homology based rules					
	1	2	3	1		2		3	
				M	N	M	N	M	N
SEQ	6	6	5	83	100	67	50	40	40
SIM	12	13	15	83	75	62	46	40	33
STR	4	2	7	75	75	50	50	0	43
SEQ + SIM	9	11	13	89	78	64	64	31	31
SEQ + STR	5	5	4	60	60	60	80	50	60
SIM + STR	11	13	15	82	64	69	46	27	20
SEQ + SIM + STR	13	13	13	70	70	77	38	23	23

M is the number of rules predicting more than one homology class. A rule predicts more than one homology class if there is more than one sequence similarity cluster in the correct test predictions. N is the number of rules predicting new homology classes. A rule predicts a new homology class if there is a sequence similarity cluster in the test predictions that has no members in the training data.

The number of rules found for each level and sequence description type are given in Table 5. At the top level (1) the order of frequency of rules found for the different classes was as follows: 'Cell processes', 'Macromolecule metabolism', 'Structural elements', 'Extrachromosomal', and 'Global functions'. This is the same order as the frequency of these classes. However, it is interesting that no rules were found for the class 'Metabolism of small molecules' which is the most frequent level 1 class (40% of all classified ORFs belong to it). This may be caused by the heterogeneity of the class or a machine learning artefact of its high frequency—rules were found for the class 'small-molecule metabolism' in *M.tuberculosis* (King *et al.*, 2000a). At level 2 the most common classes for rules were the following: 'Transport/binding proteins', 'Laterally acquired elements', 'Energy metabolism carbon', 'Ribosome constituents', 'Degradation of small molecules', 'Cell envelope', and 'Amino acid biosynthesis'. The frequent occurrence of rules for the class 'Transport/binding proteins' can be explained by it being the most common level 2 class for ORFs, and by transport proteins having relatively easily identified patterns of hydrophobicity. However, note that there exist many sequences covered by these rules that have not been predicted using SIM. The frequency of rules for the class 'Laterally acquired elements' can be explained by the typically different distribution of residues in these proteins; this class has only ∼5% of all classified ORFs. A similar explanation probably applies to 'Ribosome constituents' (see also King *et al.*, 2000b) and the class 'Cell envelope'. 'Energy metabolism carbon' is the third most common level 2 class. It is interesting that many rules were found for this class, but not for

SEQ Level 2 Rule 56

IF

the estimated pI of the ORF is > 9.11 AND

the amino_acid_pairs_qp <= 1 AND

the amino_acid_pairs_vg > 2 AND

THEN the class is "Transport/binding proteins"

Fig. 2. This rule had an accuracy of 64.5% on the test set (20/31), the default accuracy was (21%). The estimated probability of this test accuracy occurring by chance is $\sim 2.4 \times 10^{-7}$. Examination of the 11 test set 'errors' shows that 8 of them (b3633, b2288, b1590, b2056, b2484, b3140, b4315, b2610) are membrane proteins with arguable transport/binding functions which illustrates the ambiguity involved in assigning proteins to functional classes. Applying the rule to the ORFs of unknown function gives 49 predictions (2% of unassigned ORFs). It is unclear why the dipeptide qp should disfavour 'Transport/binding proteins' and the dipeptide vg should favour them. The correctly predicted proteins in the test set are [(b2771, MFS family of transport protein, 3rd module, function unknown), (b3093, exuT, MFS family of transport protein transport of hexuronates, 2nd module), (b2182, bcr, MFS family of transport protein bicyclomycin resistance protein, transmembrane protein, 2nd module), (b2587, kgtP, MFS family of transport protein α-ketoglutarate permease, 1st module), (b1528, ydeA, ABC superfamily, membrane, putative membrane component of ABC transport system appears to facilitate arabinose export contributes to control of arabinose regulon), (b3270, yhdY, ABC superfamily, membrane, paral putative membrane component of transport system), (b0198, yaeE, ABC superfamily, membrane, paral putative membrane component of transport system), (b2546, ABC superfamily, membrane, paral putative membrane component of ABC transport system, 2nd module), (b1496, yddA, ABC superfamily, atp_bind, paral putative ATP-binding module, 2nd module), (b4067, yjcG, SSS family transport protein), (b3258, panF, SSS family transport protein sodium/pantothenate symporter, 1st module), (b1336, ydaH, ArAAP family *p*-aminobenzoyl-glutamate utilization paral putative pump protein, transport, 1st module), (b1907, tyrP, ArAAP family tyrosine-specific transport system), (b0770, DASS family of transport protein), (b0341, cynX, cyanate transport), (b2987, pitB, PiT family low-affinity phosphate transport, 1st module), (b0591, ybdA, paral putative POT family of transport protein, 1st module), (b1663, ydhE, MATE family of transport protein, 2nd module), (b0336, codB, NCS1 family transport protein cytosine permease/transport, 2nd module), (b3907, rhaT, GntP family rhamnose permease, L-rhamnose-H+ symporter membrane protein, 1st module)]. The classification errors were [(b3633, kdtA, 3-deoxy-D-manno-octulosonic-acid transferase, KDO transferase), (b2288, nuoA, NADH dehydrogenase I chain A), (b1590, putative DMSO reductase anchor subunit), (b2056, putative colanic acid polymerase), (b2484, hydrogenase four membrane subunit (1st module), (b3140, agaD, PTS system *N*-acetylglucosamine enzyme IID component 1), (b4315, fimI, fimbrial protein internal segment), (b2610, ffh 4.5S-RNP protein, 1st module GTPase activity), (b3320, rplC, 50S ribosomal subunit protein L3), (b3984, plA, 50S ribosomal subunit protein L1 regulates synthesis of L1 and L11), (b3681, glvG, probable 6-phospho-beta-glucosidase)].

SIM level 2 Rule 23

IF

a homologous protein was found in an Arthopda sp. AND

a homologous protein was not found in *Helicobacter pylori* with molecular weight > 55,220 Daltons AND

a homologous protein was found in a Helicobacter group sp. with molecular weight > 55,220 Daltons AND

a homologous protein was not found in a Embryophyta sp. with the keyword "repeat".

THEN the class is "Energy metabolism carbon"

Fig. 3. This rule had an accuracy of 75% on the test set (3/4), the default accuracy was 9.8%. The probability of this happening by chance is estimated at $\sim 3.5 \times 10^{-3}$). The correctly predicted proteins in the test set are (b1468, narZ 'nitrate reductase 2 alpha subunit, 1st module), (b2283, nuoG, NADH dehydrogenase I chain G), (b2206, napA, periplasmic nitrate reductase in complex with NapB, 1st module). The 'error' is (b1872, bisZ, biotin sulfoxide reductase 2, 1st module) which is classed as 'Central intermediary metabolism'. We would argue that this protein is as much part of 'Energy metabolism carbon' as the other proteins correctly classified as such. The rule predicts the function of four ORFs of unassigned function. The Embryophyta include almost all multicellular land plants. The causative mechanism of this rule is obscure.

'Macromolecule synthesis, modification'—the second most frequent class. This implies that the proteins involved in 'Energy metabolism carbon' are more homogeneous than those in 'Macromolecule synthesis, modification'. At level 3, the most specific level, the most common classes for rules were: 'Transposon-related functions', 'Ribosomal proteins', 'MFS family', 'Surface structures', 'Global regulatory functions', 'ABC superfamily (membrane)', 'ABC superfamily (atp_bind)', and 'Chemotaxis and mobility'. These results are similar to the results at level 2. Few rules are found for classes involving enzymes. This may be because most such classes have very few ORFs, making it difficult to generalise.

Four typical rules are selected to illustrate the forms of rule formed (Figures 2–5), one rule each from the descriptor types (SEQ, SIM, STR, and SEQ + SIM + STR). To allow direct comparison between the descriptor types, all the rules except that from SEQ are for the class 'Energy metabolism carbon' (no rule was found for this class with SEQ). Each of the four rules is not exclusively based on sequence homology (see below). The rules perform well on the test data and predict a number of ORFs of unknown function. Analysis of the errors made by the prediction rules makes it clear that many of them are not really errors and are instead artefacts of the annotation process—they are biologically justifiable. The main cause of these annotation errors is that existing functional hierarchies often give only one function per

STR Level 2 Rule 50
IF
two α-helices are not predicted i, i+2, the first ≤1, and the second >1 AND
three β-strand predictions are predicted j, j+2, j+4, the first ≤6, the second ≤830, and the third ≤6 AND
three coil predictions are not predicted k, k+2, k+4, the first >5, the second >3, and the third >10 AND
three coil predictions are predicted l, l+2, l+4, the first >10, the second ≤6, and the third ≤6 AND
three coil predictions are predicted m, m+2, m+4, the first >11, the second ≤830, and the third ≤10
THEN the class is "Energy metabolism carbon"

Fig. 4. This rule had an accuracy of 50% on the test set (3/6), the default accuracy was (9.8%). The probability of this happening by chance is estimated at $\sim 1.5 \times 10^{-2}$). The correctly predicted proteins in the test set are [(b4071, nrfB, formate-dependent nitrite reductase, a penta-haeme cytochrome c), (b2720, hycF, probable iron–sulfur protein of hydrogenase 3, part of formate hydrogenlyase complex), (b2724, hycB, probable small subunit of hydrogenase-3 iron–sulfur protein, part of formate hydrogenlyase complex)]. The errors are [(b3721, bglB, phospho-β-glucosidase B, cryptic), (b1412, acpD, acyl carrier protein phosphodiesterase'), (b2587, kgtP, MFS family of transport protein α-ketoglutarate permease, 1st module)], all of which are related to 'Energy metabolism carbon'. The structural meaning of the rule is unclear. The rule predicts the function of 18 ORFs of unassigned function.

SEQ + SIM + STR Level 2 Rule 20
IF
two β-strands are predicted at i, i+2, the first >3, and the second >1 AND
a homologous protein was found in a Kinetoplastida sp. with the keyword "transmembrane" AND
a homologous protein was not found in a Epsilon subdivision sp. with the keyword "inner_membrane"
THEN the class is "Energy metabolism carbon"

Fig. 5. This rule had an accuracy of 50% on the test set (6/12), the default accuracy was (9.8%). The probability of this happening by chance is estimated at $\sim 5 \times 10^{-4}$). The correctly predicted proteins in the test set are [(b2283, nuoG, NADH dehydrogenase I chain G), (b2484, hydrogenase 4 membrane subunit, 1st module), (b2243, glpC, sn-glycerol-3-phosphate dehydrogenase, anaerobic, K-small subunit, 2nd module), (b4379, yjjW, pyruvate formate lyase activating enzyme), (b2720, hycF, probable iron–sulfur protein of hydrogenase 3, part of formate hydrogenlyase complex), (b2724, hycB, probable small subunit of hydrogenase-3 iron–sulfur protein, part of formate hydrogenlyase complex)]. The errors are [(b4359, mdoB, phosphoglycerol transferase I, add phosphoglycerols to OPG backbone), (b3385, gph, phosphoglycolate phosphatase), (b2316, accD, acetylCoA carboxylase carboxytransferase component β subunit), (b2056, putative colanic acid polymerase), (b3653, gltS, GltS family glutamate transport, 2nd module), (b3469, P-type ATPase family zinc-transporting ATPase, 2nd module)]. Half of these 'errors' (mdoB, gph, and accD) are clearly related to the functional class 'Energy metabolism carbon'. The rule predicts the function of 40 ORFs of unassigned function (~2%).

gene (Kell and King, 2000). This means that the error rates of the DMP rules on the test data are pessimistically biased, i.e. their true accuracy is probably higher than estimated in the tables.

The number of ORFs of unknown functions predicted by the rules from the different types of sequence description are given in Table 4. The sequence description method with the highest coverage WTD_VOTE_ALL predicts 40% of the unassigned ORFs. Given that *E.coli* is a very well studied organism, this is a notable extension of its annotation. A lower percentage of unknown ORFs are predicted compared to the test set of known function. This is expected as the distribution of sequences of ORFs of unknown function is likely to be different from those of assigned function.

For those proteins correctly predicted by each rule we carried out all-against-all PSI-BLAST searches. If all the proteins could be linked together by PSI-BLAST scores <10 then the proteins were considered homologous (note that this is a very liberal definition). It was found that many of the predictive rules were more general than possible using sequence homology. This was shown in two ways: the rules correctly predict the function of sets of proteins that are not homologous to each other, and they correctly predict the function of proteins that are not homologous to any in the training data (Table 5). A larger percentage of non-homology rules were identified at level 1 (60–90%) compared with level 3 (30–60%). Level 2 was intermediate. This was expected as the lower level functional classes are more homogeneous and homologous sets of proteins within these classes make a larger percentage of them. The three basic types of descriptor (SEQ SIM STR) all seem about equal in their ability to describe non-homology based rules.

DISCUSSION AND CONCLUSION

DMP rules provide a novel and powerful way of predicting an ORFs function from its sequence. Although they cannot provide the specificity of the predictions provided by standard SIM searches, the DMP approach has the ability to make correct predictions when standard methods fail, and to provide independent conformation of a prediction made by standard methods. As such they are an effective new technique in the bioinformatician's tool-box.

Perhaps the most interesting biological feature of DMP rules is that many of them can predict function in the absence of detectable sequence homology. The existence of such rules was unforeseen—given the notoriously com-

plicated mappings between function and structure, and structure and sequence. Possible causative mechanisms of such DMP rules are: detection of remote homologous sequences too distant as to be undetectable by sequence analysis (Henikoff *et al.*, 1997; Tatusov *et al.*, 1997); convergent evolution forcing proteins with similar function to resemble each other; and horizontal evolution transferring functional related groups of protein into the organisms.

The DMP rules based on homology are also interesting as they provide a novel way of detecting homology. We have demonstrated in this paper, and in King *et al.* (2000a,b) that the DMP rules correctly predict the function of proteins missed in the original annotations. Now it could be argued that these functions could have been detected using a less conservative setting of the probability of SIM match (i.e. higher e value threshold). It is hard to disprove this claim, as the original annotation lacks the required information, and the current absence of a cross-species functional hierarchies and ontologies (see below) makes automated annotation difficult. However, it is likely that combined with standard homology detection programs the DMP rules could accurately predict more distant homologies than existing methods alone. This is because the DMP rules use a different bias (in machine learning terms: Mitchell, 1997) from standard approaches. Combining prediction methods with different biases is a standard method of improving prediction method accuracies (Dietterich, 1997).

The inference of homology based on SIM is generally based on a threshold approach: homology is inferred if a SIM search detects a match over a threshold probability; if the match is below this threshold, no matter by how little, no homology is inferred. This is a mistake. In decision theory this approach is equivalent to assigning a particular loss function to the errors: when assigning homology there are two types of error possible: errors of commission (where an ORF is predicted to have a function that it does not have), and errors of omission (where an ORF is not predicted to have a function that it does have). The costs of making these two types of error are not necessarily equal, and depend on the relative cost of investigating an erroneous prediction compared to missing a biologically interesting correct prediction. Generally we wish to make the decision which minimises the expected loss, and this is achieved if:

(probability(homologous|sequence,background data)/
probability(not homologous|sequence, background data))
$>$ (loss(error of commission)/loss(error of omission)).

This equation makes explicit why annotators generally use high thresholds of probability to assign homology—i.e. they are very conservative. They intuitively assume that cost of making errors of commission are more costly than errors of omission. In this they are generally correct, as assignments of function are generally given without probability values, and so errors may propagate without control through the databases. However, we do not all share the same loss function as the annotators. In particular, if your loss function has a lower cost on errors of commission than the annotators, then the annotation is useless, as the annotators have thrown away the information in the SIM search. If you have a higher cost on errors of commission then the assignment is non-optimal as you will make more errors of commission than you could afford. DMP rules can be tuned to find rules with different loss functions. This can be done both by changing the selection criteria in the validation set, and by choosing different sequence descriptor types to give different coverages and averages (Tables 4). It is possible to show that certain prediction methods 'dominate' others over a large range of reasonable loss functions (Provost and Fawcett, 1997).

The three basic types of description (SEQ, SIM and STR) can be improved upon. Although SEQ is surprisingly effective, it fails to include all the explicit sequential information in a protein sequence. The following residue sequences would have exactly the same SEQ description: [aacaacaa], [acaaaaca]. This is clearly not ideal. Possible avenues for improvement are to use wavelets (Mallat, 1989), the Santa Cruz approach (Jaakkola *et al.*, 1999), and the direct use of ILP data mining. Sequences are inherently relational and poorly described using attribute vectors, and Warmr could directly include relational information on sequence and be used to find frequent sub-sequences that characterise sequences. The most obvious way to improve SIM would be to include multiple sequence alignment information. In STR information could be included about: the relation between predicted secondary structure elements, the distribution of secondary structure types (mostly α, mostly β), probabilities, etc. Another approach to improving would be to involve fold prediction, this would allow direct linkage of existing knowledge of structure to function. For this to work fold prediction does not necessarily need to be very accurate, only informative.

Our approach to DMP has so far been organism specific. This has been because of the widely different biology in the species studied (*M.tuberculosis, E.coli* Kell and King, 2000), and by the lack of a consistent cross-species functional hierarchy. The only existing hierarchy that extends across species is the EC enzyme classification system. However this is far from ideal, as it is restricted to enzymes and is more chemically based than biologically. Despite this we would expect from our results that DMP would be capable of predicting enzyme classifications. Work has started on developing cross-species functional hierarchies with the formation of controlled vocabularies and ontologies (The Gene Ontology Consortium, 2000;

http://www.geneontology.org/ Rison *et al.*, 2000). When such ontological work has annotated a sufficient number of species it will be possible to search for pan-specific rules relating sequence to function.

REFERENCES

Aha,D., Kibler,D. and Albert,M. (1991) Instance-based learning algorithms. *Mach. Learn.*, **6**, 37–66.

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

Bairoch,R. and Apweiler,A. (1999) The SWISSPROT protein sequence data bank and its supplement TrEMBL in 1999. *Nucleic Acids Res.*, **27**, 49.

Bauer,E. and Kohavi,R. (1999) An empirical comparison of voting classification algorithms: bagging, boosting, and variants. *Mach. Learn.*, **36**, 105–139.

Blattner,F.R., Plunkett,G., Bloch,C.A., Perna,N.T., Burland,V., Riley,M., Collado-Vides,J., Glasner,J.D., Rode,C.K., Mayhew,G.F., Gregor,J., Davis,N.W., Kirkpatrick,H.A., Goeden,M.A., Rose,D.J., Mau,B. and Shao,Y. (1997) The complete genome sequence of *Escherichia coli* K-12. *Science*, **277**, 1453–1461.

Brenner,E. (1999) Errors in gene annotation. *Trends Genet.*, **15**, 132–133.

Dehaspe,L., Toivonen,H. and King,R.D. (1998) Finding frequent substructures in chemical compounds. In Agrawl,R., Stolorez,P. and Piatetsky-Shapiro,G. (eds), *Proceedings of the Fourth International Conference on Knowledge Discovery and Data Mining.* AAAI Press, Menlo Park, CA, pp. 30–36.

Dietterich,T.G. (1997) Machine learning research: four current directions. *AI Magazine*, **18**, 97–136.

Duda,R. and Hart,P. (1973) *Pattern Classification and Scene Analysis*. Wiley, New York.

Fayyad,U., Piatetsky-Shapiro,G., Smyth,P. and Uthurusamy,R. (1996) *Advances in Knowledge Discovery and Data Mining*. AAAI/MIT Press, Boston, MA.

Henikoff,S., Greene,E.A., Pietrokovski,S., Bork,P., Attwood,T.K. and Hood,L. (1997) Gene families: the taxonomy of protein paralogs and chimeras. *Science*, **278**, 609–614.

Jaakkola,T., Diekhans,M. and Haussler,D. (1999) Using the Fischer kernal method to detect remote sequence homologies. *Seventh International Conference on Intelligent Systems for Molecular Biology.* AAAI Press, New York, CA, pp. 149–155.

des Jardins,M., Karp,P.D., Krummenacker,M., Lee,T.J. and Ouzounis,C.A. (1997) Prediction of enzyme classification from protein sequence without the use of sequence similarity. *Proceedings of the 5th International Conference on Intelligent Systems for Molecular Biology*, 93–99.

Jones,D.T., Taylor,W.R. and Thornton,J.M. (1992) A new approach to protein fold recognition. *Nature*, **358**, 86–89.

Kelly,L.A., MacCallum,R.M. and Sternberg,M.J.E. (2000) Enhanced genome annotation using structural profiles in the program 3D-PSSM. *J. Mol. Biol.*, **299**, 501–522.

Kell,D.B. and King,R.D. (2000) On the optimization of classes for the assignment of unidentified reading frames in functional genomics programmes: the need for machine learning. *Trends Biotechnol.*, **18**, 93–98.

King,R.D., Karwath,A., Clare,A. and Dehapse,L. (2000a) Genome scale prediction of protein functional class from sequence using data mining. In Ramakrishnan,R., Stolfo,S. and Bayardo,R. (eds), *The Sixth ACM SIGKDD International Conference on Knowledge Discovery and Data Mining.* The Association for Computing Machinery, New York, pp. 384–389.

King,R.D., Karwath,A., Clare,A. and Dehapse,L. (2000b) Accurate prediction of protein functional class in the *M.tuberculosis* and *E.coli* genomes using data mining. *Yeast (Comparative and Functional Genomics)*, **17**, 283–293.

Lavrac,N. and Dzeroski,S. (1994) *Inductive Logic Programming: Techniques and Applications*. Ellis Horwood, Chichester, UK.

Mallat,S.G. (1989) A theory for multiresolution signal decomposition and wavelet representation. *IEEE Trans. Patt. Anal. Mach. Intel.*, **11**, 674–693.

Marcotte,M., Pellegrine,M., Thompson,M.J., Yeates,T.O. and Eisenberg,D. (1999) A combined algorithm for genome wide prediction of protein function. *Nature*, **402**, 83–86.

Mitchell,T.M. (1997) *Machine Learning*. McGraw-Hill, New York.

Muggleton,S. (1991) Inductive logic programming. *New Gener. Comput.*, **8**, 295–318.

Munakata,T. (1999) Knowledge discovery. *Comm. ACM*, **41**, 26–29.

Ouali,M. and King,R.D. (2000) Cascaded multiple classifiers for secondary structure prediction. *Protein Sci.*, **9**, 1162–1176.

Pearson,W.R. and Lipman,D.J. (1988) Improved tools for biological sequence comparison. *Proc. Natl Acad. Sci. USA*, **85**, 2444–2448.

Piatetsky-Shapiro,G. and Frawley,W. (1991) *Knowledge Discovery in Databases*. MIT Press, Boston, MA.

Provost,F. and Fawcett,T. (1997) Analysis and visualization of classifier performance: comparison under imprecise class and cost distributions. In Heckerman,D., Mannila,H. and Pregibon,D. (eds), *Proceedings of KDD-97.* AAAI Press, Menlo Park, CA, pp. 43–48.

Quinlan,R. (1993) *C4.5: Programs for Machine Learning*. Morgan Kaufmann, San Mateo, CA.

Riley,M. and Labedan,B. (1996) *E.coli* gene products: physiological functions and common ancestries. In Neidhardt,F. *et al.* (eds), Escherichia coli *and* Salmonella*: Cellular and Molecular Biology,* pp. 2118–2200.

Rison,S.C.G., Hodgman,T.C. and Thornton,J.C. (2000) Comparison of functional annotation schemes for genomes. *Funct. Integr. Genomics*, **1**, 56–69.

Stawiski,E.W., Baucom,A.E., Lohr,S.C. and Gregoret,L.M. (2000) Predicting protein function from structure: unique structural features of proteases. *Proc. Natl Acad. Sci. USA*, **97**, 3954–3958.

Tatusov,R.L., Koonin,E.V. and Lipman,D.J.A. (1997) Genomic perspective on protein families. *Science*, **278**, 631–637.

TB_gene_list, http://www.sanger.ac.uk/Projects/M_tuberculosis/Gene_List/.

The Gene Ontology Consortium, (2000) Gene ontology: tool for the unification of biology. *Nature Genet.*, **25**, 25–29.

Ullman,J.D. (1988) *Principles of Databases and Knowledge-base Systems*. Vol. 1, Computer Science Press, Rockville, MD.

Wootton,J.C. and Federhen,S. (1993) Statistics of local complexity in amino acid sequences and sequence databases. *Comput. Chem.*, **17**, 149–163.