Genome Scale Prediction of Protein Functional Class from Sequence using Data Mining

Ross D. King
Department of Computer Science
University of Wales, Aberystwyth
Penglais, Aberystwyth
SY23 3DB, Wales, U.K
+44 1970 622432
rdk@aber.ac.uk

Andreas Karwath
Department of Computer Science
University of Wales, Aberystwyth
Penglais, Aberystwyth
SY23 3DB, Wales, U.K
+44 1970 621922
aak97@aber.ac.uk

Amanda Clare
Department of Computer Science
University of Wales, Aberystwyth
Penglais, Aberystwyth
SY23 3DB, Wales, U.K
+44 1970 621922
ajc99@aber.ac.uk

Luc Dehaspe
Department of Computer Science, Katholieke Universiteit Leuven
PharmaDM
Celestijnenlaan 200A,
B-3001, Belgium
++32 16 327643
Luc.Dehaspe@PharmaDM.com

ABSTRACT

The ability to predict protein function from amino acid sequence is a central research goal of molecular biology. Such a capability would greatly aid the biological interpretation of the genomic data and accelerate its medical exploitation. For the existing sequenced genomes function can be assigned to typically only between 40-60% of the genes [4,8,12,7]. The new science of functional genomics is dedicated to discovering the function of these genes, and to further detailing gene function [10,27,17,6]. Here we present a novel data-mining [24,18] approach to predicting protein functional class from sequence. We demonstrate the effectiveness of this approach on the *Mycobacterium tuberculosis* [8] genome. Biologically interpretable rules are identified that can predict protein function even in the absence of identifiable sequence homology. These rules predict 65% of the genes with no previous assigned function in *Mycobacterium tuberculosis* (the bacteria which causes TB) with an estimated accuracy of 60-80% (depending on the level of functional assignment). The rules give insight into the evolutionary history of the organism.

Categories and Subject

Database Applications, Learning, Life and Medical Sciences

General Terms

Data mining, Concept learning, Biology and genetics.

1. INTRODUCTION

The formation of a theory to explain a set of observations is central to science. Computer based methods to assist in this process are becoming increasingly important [20]. Such methods are especially needed in molecular biology, where there is an overwhelming flood of new data. Here we demonstrate the effectiveness of automatic scientific discovery on an important scientific problem. We successfully apply a novel data mining approach to the problem of predicting protein functional class from sequence.

To predict the biological functional class of proteins directly from sequence what is abstractly required is a discrimination function [10] that maps sequence to biological function.

To predict protein function directly from sequence what is abstractly required is a discrimination function10 that maps sequence to biological function. The existing sequence homology recognition methods can be viewed as examples of such functions: methods based on direct sequence similarity [23,2] can be considered as nearest neighbour type functions [11] (in sequence space), and the more complicated homology recognition methods based on motifs/profiles [28] resemble case-based learning functions [1]. The creation of annotated databases of protein function has now opened up the possibility of automatically identifying more general forms of discrimination function using data.

2. DATA

For analysis, we selected the tuberculosis genome, probably the prokaryote genome of greatest medical importance. According to the World Health Organization (WHO), tuberculosis kills 2 million people each year. Their concern about the growing epidemic has led the WHO to declare tuberculosis a global emergency [15].

We used 3,924 genes [8] * (over 4 million base pairs) with functional class assignments from the Sanger Centre [14]. (Note that there are errors in annotation of function [5], and this adds "noise" to the data mining process [18]). The assignments of function are organised in a strict hierarchy (tree), where each higher level in the tree is more general than the level below it, and

KDD 2000, Boston, MA USA
© ACM 2000 1-58113-233-6/00/08 ...$5.00

[*] For readability reasons we used "gene" throughout the paper, knowing that "potenial gene" or open reading frame (ORF) should be used.

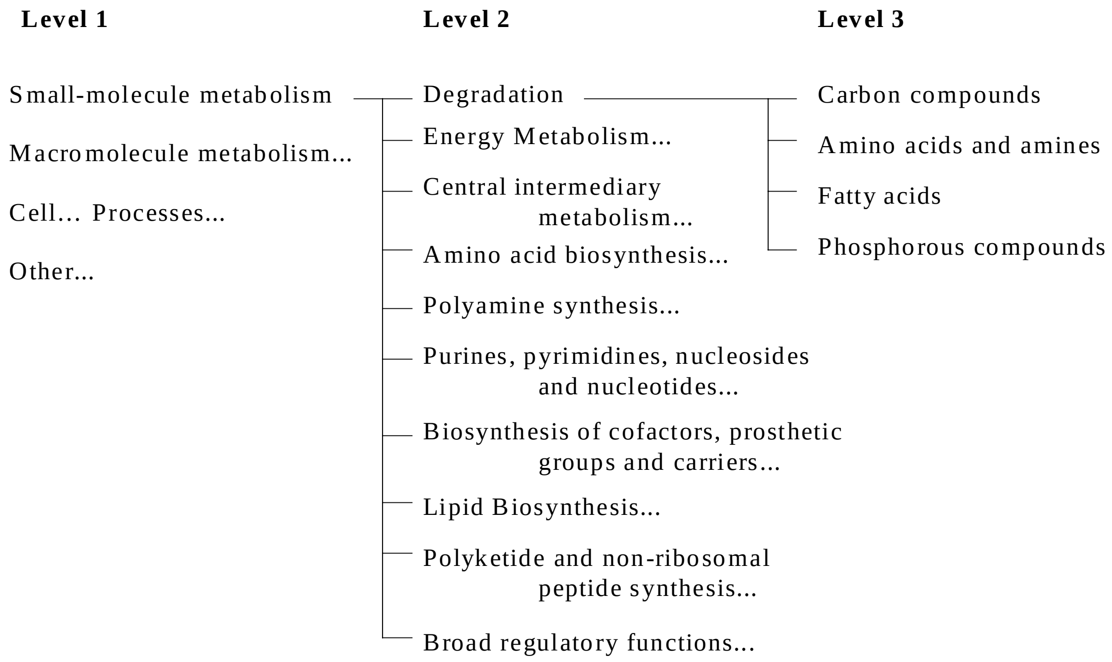
Level 1
Level 2
Level 3
Small-molecule metabolism
Macromolecule metabolism...
Cell... Processes...
Other...
Degradation
Energy Metabolism...
Central intermediary metabolism...
Amino acid biosynthesis...
Polyamine synthesis...
Purines, pyrimidines, nucleosides and nucleotides...
Biosynthesis of cofactors, prosthetic groups and carriers...
Lipid Biosynthesis...
Polyketide and non-ribosomal peptide synthesis...
Broad regulatory functions...
Carbon compounds
Amino acids and amines
Fatty acids
Phosphorous compounds

Figure 1. An example subset of the genes functional hierarchy. This example has only three out of four possible classification levels.

the leaf nodes are the individual functions of proteins. A subsection of the function hierarchy is shown in figure 1.

For example, a typical protein in the tubercle bacillus is L-fuculose phosphate aldolase (Rv0727c fucA), its top-level class assignment is “Small-molecule metabolism”, its second-level class is “Degradation”, and its third-level class is “Carbon compounds”. We attempted to learn discriminatory functions for every level of the functional hierarchy. Success on these different levels would demonstrate the generality of the approach.

To generate the database to mine we formed a single deductive database of genes and their known functional assignments. We then processed this data to form sequence descriptions of the genes. Therefore, these descriptions are solely based on features that can be computed from sequence alone. The most commonly used technique to gain information about a sequence is to run a sequence similarity search, and this was used as the starting point in forming descriptions. The basic data structure in the deductive database is the result of a PSI-BLAST search [2] (we used the parameters: e = 10, h = 0.0005, j = 20, NRProt 16/11/98). NRProt is a protein sequence database merging together protein sequences with less than 100 percent sequence identity from a variety of multi-genome protein databases. Using each gene, and each protein identified as having sequence similarity to it, we formed an expressive description based on: the frequency of singlets and pairs of residues in the gene; the phylogeny (“family tree”) of the organism from which each protein was obtained - from SWISS-PROT [3] (a standard protein database); SWISS-PROT protein keywords (membrane, transmembrane, inner_membrane, outer_membrane, repeat, plasmid, and alternative_splicing); and the length and molecular weight of the gene. This description resembles a “phylogenic profile” [17], but is more general and expressive. In total 5,895,649 facts were generated. Table 2 shows the available database facts and their description.

3. DATA MINING METHOD

We then mined this database to generate rules that predict protein functional class from sequence description. This was done using a combination of clustering and rule learning (see Figure 2).

This hybrid approach has proved successful in the past on other scientific discovery tasks [9]. It is powerful because clustering improves the representation for learning (using the expressive power of inductive logic programming – ILP [19]), and discrimination efficiently exploits the pre-labeled examples. WARMR [9] is an ILP data mining algorithm that is used to identify frequent patterns (conjunctive queries) in the sequence descriptions. In this experiment roughly 18,000 frequent queries were discovered. These were converted into 18,000 Boolean attributes for rule learning, where an attribute gets value 1 for a specific gene if the corresponding query succeeds for that gene. Conversely, if the query fails, the corresponding attribute is assigned value 0.

The machine learning algorithms C4.5 and C5 [26] were used to induce rules that predict function from the attributes. Good rules were selected on a validation set, and the unbiased accuracy of these rules estimated on a test set. Rules were selected to balance accuracy with unidentified gene coverage. In general the correct balance of accuracy and coverage for any particular application depends on the relative cost of making errors of commission and omission [25] (making incorrect predictions v missing genes). The system can be tuned to select

Table 1. Database facts and their description. These facts are generated for each of the genes.

Database argument	Description
hom(A)	refers to a homologous protein (*A*) found by PSI-BLAST.
keyword(A, Word)	refers to a SwissProt keyword found in *A*.
classification(A, Class)	refers to the phylogenic classification of the organism *A* came from, taken from SwissProt.
species(A, Species)	refers to the species of *A*, taken from SwissProt.
mol_wt_rule(A, Weight)	refers to the molecular weight of *A*: 1 very low, 2 low, 3 medium, 4 high, and 5 very high.
amino_acid_ratio_rule(A, Residue, Weight)	refers to the percentage composition of the residue in the sequence.
e_val_rule(A, Weight)	refers to the PSI-Blast sequence similarity measure (note that a low value means a high sequence similarity).

e_val_gt(Value)
e_val_lteq(Value)
	refers to the PSI-Blast sequence similarity measure, greater than or less than/equal to a certain value

mol_wt_lteq(A, Weight)
mol_wt_gt(A, Weight)
	refers to the molecular weight of *A* being greater than or less than/equal to some value

amino_acid_pairs_wg()	and others similar, refers to the number of pairs of these two amino acids, in this case tryptophan and glycine
amino_acid_pair_ratio_qh (Ratio)	and others similar, refers to the ratio of one amino acid to another in the gene, in this case the ratio of glutamine(q) to histidine(h). This ratio is not a percentage, not out of a hundred, instead it's a ratio out of a thousand. So for example 2.8 means 0.28%.
amino_acid_ratio_g (Percentage)	and others similar, refers to the percentage composition of the residue in the sequence of the gene, in this case the percentage of glycine

psi_iter_gt(Number)
psi_iter_lteq(Number)
	refers to the number of iterations of the PSI_BLAST search (greater than or less than/equal to some number)

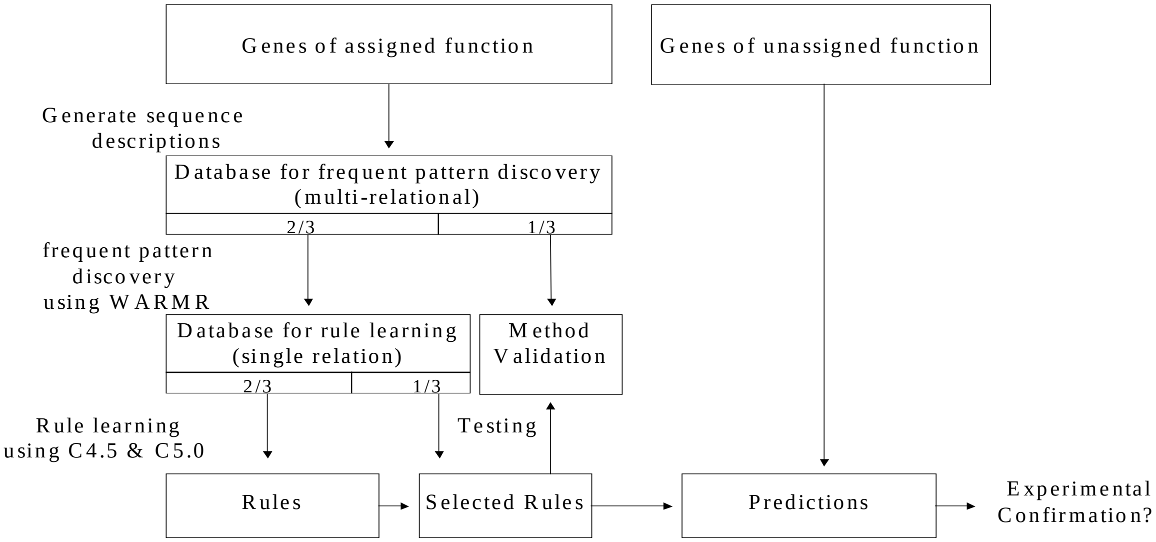
Genes of assigned function
Genes of unassigned function
Generate sequence descriptions
Database for frequent pattern discovery (multi-relational)
2/3
1/3
frequent pattern discovery using WARMR
Database for rule learning (single relation)
2/3
1/3
Method Validation
Rule learning using C4.5 & C5.0
Testing
Rules
Selected Rules
Predictions
Experimental Confirmation?

Figure 2. Flow chart of the data mining methodology. The genes of known functions were split randomly, one third to be held out for the final testing and two thirds to be used to generate prediction rules. The data used to generate the rules were in turn split randomly, two thirds to actually generate the rules and one third to be used as validation set for selecting good rules according to accuracy and coverage. After the selection of good rules, we tested their accuracy on the held-out test data, and also used them to predict biological function for the genes of currently unknown function (classified as "Unkown" or "Conserved Hypotheticals").

different balances. The prediction rules were then applied to genes that have not been assigned a function.

4. RESULTS

It was possible to find good rules that predict function from sequence at all levels of the functional hierarchies, as shown in Table 2.

The test accuracy of these rules is far higher than possible by chance. Of the genes originally in the “Conserved Hypothetical” or “Unknown” function classes, 985 (65%) were predicted to have a function at one or more levels of the hierarchy. The rule learning data, the rules, and the predictions, are given at: http://www.aber.ac.uk/~dcswww/Research/bio/ProteinFunction/.

We illustrate the value of the rules by describing rule TB_C50_1_26 shown in figure 3.

If

the percentage composition of lysine in the gene is > 6.6%

Then

the ORF has the functional class “Macromolecule metabolism”

Figure 3. Example rule found for the Macromolecule metabolism functional class.

Table 2. The number of rules found are those selected on the validation set. A rule predicts more than one homology class if there is more than one sequence similarity cluster in the correct test predictions. 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. Average test accuracy is the accuracy of the predictions on the test proteins of assigned function (if conflicts occur, the prediction with the highest *a priori* probability is chosen). Default test accuracy is the accuracy that could be achieved by always selecting the most populous class. “New functions assigned” is the number of genes of unassigned function predicted.

	Level 1	*Level 2*	*Level 3*	*Level 4*
Number of rules found	25	30	20	3
Rules predicting more than one homology class	19	18	8	1
Rules predicting a new homology class	14	15	1	0
Average test accuracy	62%	65%	62%	76%
Default test accuracy	48%	14%	6%	2%
New functions assigned	886 (58%)	507 (33%)	60 (4%)	19 (1%)

This top-level rule is 85% (11/13) accurate on the test set (the probability of this result occurring by chance is estimated at 1.2x10^{-5} as the class Macromolecule metabolism covers ~25% of examples). The rule correctly predicts the following proteins (rpsG (S7), rpsI (S9), rpsL (S12), rpsT (S20), rplJ (L10), rplP (L16), rplS (L19), rplX (L24), rpmE (L31), rpmJ (L36), infC (IF-3)). These proteins are all involved in protein translation. When the training data are included the rule covers 46 out of the 58 proteins known to be involved in ribosomal protein synthesis and modification. The two errors (of commission) made in the test data were groEL2 a “60 kD Chaperonin 2 gene” and Rv3583c a “putative transcriptional regulator”. The rule predicts the function of five genes classed as “Conserved Hypotheticals” (Rv566, Rv854, Rv910, Rv2185, Rv2708) and ten genes classed as “Unknowns” (Rv123, Rv810, Rv909, Rv1893, Rv1955, Rv2061, Rv2517, Rv2819, Rv2822, Rv3718). The prediction rule is consistent with protein chemistry, as lysine is positively charged which is desirable for interaction with negatively charged RNA. The choice of lysine over arginine for the positively charged residue may be connected with the high GC content of the M. tuberculosis genome [8] - lysine is coded by the codons AAA and AAG; while arg is coded by CGU, CGC, CGA, CGG; and his by CAT and CAC.

If
there exists a homologous protein in SwissProt with the keyword "membrane" and
there exists a homologous protein in *Bacillus subtilis*
and there does not exist a homologous protein with very low molecular weight, a large percentage of glutamic acid, and medium sequence similarity
and there does not exist a homologous protein in SwissProt with good sequence similarity, low percentage of cysteine, the keyword "transmembrane" and a fairly high molecular weight
there does not exist a firmicutes sp. protein in SwissProt with the keyword "transmembrane", with medium molecular weight, and a very high amount of low entropy sequence
and there exists a homologous mammalian protein in SwissProt with the keyword "repeat" with very high molecular weight
Then
the ORF has the functional class "Degradation of macromolecules".

Figure 4. A more complex rule for the classification of “Degradation of macromolecules”.

Not all rules are as simple as the example in figure 3, a more complex rule is shown in figure 4. This rule predicts the level two functional class “Degradation of macromolecules“. The rule is 62.5% accurate (5/8) on the test set. It predicts 3 genes, which are currently classified as “Unknown” or “Conserved Hypothetical”. The errors of commission are rplV (synthesis and modification of macromolecules), Rv1566 (Virulence) and ponA2 (penicillin binding protein).

The approach described in this paper is the first that can systematically find rules not based on homology. To show this we carried out all-against-all PSI-BLAST searches for those proteins correctly predicted by each rule. If all the proteins could be linked together by PSI-BLAST scores < 10 then the proteins were considered homologous. 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 1). Such rules provide a way of predicting function in the absence of recognisable sequence homology. The other rules, those of equal power to sequence homology, are also valuable as they provide a novel way of detecting homology.

5. DISCUSSION

The discovered rules are important in two ways: they make predictions that are useful in determining the functions of genes of currently unknown function, and they provide evolutionary insight. The actual function of a gene can only be determined by "wet" experiment. However, bioinformatic techniques such as sequence homology detection, and the prediction rules presented here, can make such experimental determination simpler. It is clearly more efficient to test a high probability hypothesis than to randomly test for possible functions. We look forward to the testing of our predictions by other workers, and we are designing automatic methods to test the rules ourselves.

The existence of general rules for predicting biological function raises the question of their evolutionary causation. How are such rules possible, given the notoriously complicated mappings between function and structure, and structure and sequence? Several possibilities exist: the rules are paralogous [16] with homology so distant as to be undetectable by sequence analysis; convergent evolution has occurred, forcing proteins with similar function to resemble each other; or horizontal evolution has transferred functional related groups of protein into the organisms. Evidence in favour of a role for distant homology is that it is possible to predict function better than random based on predicted secondary structure alone, and secondary structure is better conserved over evolution than sequence [22]. Evidence against this is that we have found little evidence for common SCOP database [21] "superfamily" and "fold" classifications for proteins predicted by the same rule. Convergent evolution seems to be the dominant factor in rules such as TB_C50_1_26 (Figure 3). Evidence for horizontal transfer of genes into M. tuberculosis is the importance of phylogeny in many rules where a paralogous explanation seems to be ruled out.

6. CONCLUSION

We have demonstrated the utility of automatic knowledge discovery techniques by showing that they can discover prediction rules that are effective and of biological interest in functional genomics. The data mining approach described is extendable to analysis of other forms of bioinformatic data, such as expression profiles, pathway analysis, structural studies, etc. [10,27,17,6]. Information from all these diverse approaches will be able to be combined together to produce more powerful predictions than any single one in isolation.

7. ACKNOWLEDGEMENTS

We would like to thank Douglas B. Kell and Michael Young from IBS and Mohammed Ouali from Computer Science, at the University of Wales, Aberystwyth; Ashwin Srinivasan of the Computing Laboratory at the University of Oxford; and Steffen Schulze-Kremer at the Resource Centre of the German Human Genome Project for helpful discussions. This work was supported by EPSRC, MRC, and British Council. Luc Dehaspe is a Fellow of the K. U. Leuven Research Council.

8. REFERENCES

[1] Aha, D., Kibler, D., Albert, M., Instance-based learning algorithms. *Machine Learning* **6**, 37-66 (1991).

[2] Altschul, S. F. *et al.*, Gapped BLAST and PSI-BLAST: a new generation of protein database search programs. *Nucl. Acid Res.*, **25**, 3389-3402 (1997).

[3] Bairoch, A., Apweiler, R., The SWISSPROT protein sequence data bank and its new supplement TrEMBL in 1999. *Nucleic Acids Res*. **27**, 49-54 (1999).

[4] Blattner, F. R. *et al.*, The complete genome sequence of *Escherichia coli* K-12. *Science* **277**, 1453-1461 (1997).

[5] Brenner, S. E., Errors in gene annotation. *Trends in Genetics* **15**, 132-133 (1999).

[6] Brown, M. P. S. *et al.*, Knowledge-based analysis of microarray gene expression data by using support vector machines. *Proc. Natl. Acad. Sci., USA* **97**, 262-267 (2000).

[7] C. elegans Sequencing Consortium, Genome Sequence of the Nematode C. elegans: A Platform for Investigating Biology *Science* **282**, 2012-2018 (1998).

[8] Cole, S. T. *et al.*, Deciphering the biology of *Mycobacterium tuberculosis* from the complete genome sequence. *Nature* **393**, 537-544 (1998).

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

[10] DeRisi, J. L., Iyer, V. R., Brown, P. O., Exploring the metabolic and genetic control of gene expression on a genomic scale. *Science* **278**, 680-686 (1997).

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

[12] Goffeau, A. *et al.*, Life with 6000 genes. *Science* **274**, 546-567 (1996).

[13] Henikoff, S., *et al*. Gene families: the taxonomy of protein paralogs and chimeras. Science **278** 609-614 (1997)

[14] http://www.sanger.ac.uk/Projects/M_tuberculosis/gene_list_full.shtm (15/03/1999).

[15] http://www.who.int/inf-fs/en/fact104.html

[16] Kell, D., King, R. D., On the optimization of classes for the assignment of unidentified reading frames in functional genomics programmes: the need for machine learning. *Trends in Biotechnology* **18**, 93-98 (2000).

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

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

[19] Muggleton, S., Inductive logic programming *New Generation Computing* **8**, 295-318 (1991).

[20] Munakata, T., Knowledge Discovery. *Comm. ACM* **41** 26-29 (1999)

[21] Murzin, A. G., Brenner, S. E., Hubbard, T., Chothia, C., SCOP: a structural classification of proteins database for the investigation of sequences and structures. *J. Mol. Biol.* **247**, 536-540 (1995).

[22] Park, J., Teichmann, S. A., Hubbard, T., Chothia, C., Intermediate Sequences Increase the Detection of Homology Between Sequences. *J. Mol. Biol.* **273**, 349-354 (1997).

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

[24] Piatetsky-Shapiro, G., Frawley W., *Knowledge Discovery in Databases* (MIT Press, 1991).

[25] Provost, F., Fawcett, T., Analysis and Visualization of Classifier Performance: Comparison under Imprecise Class and Cost Distributions. In *Proceedings of KDD-97* Heckerman, D., Mannila, H., Pregibon, D., eds. (AAAI Press, Menlo Park, 1997) pp *43-48*.

[26] Quinlan, R., C*4.5: Programs for machine learning* (Morgan Kaufmann, San Mateo, 1993).

[27] Tatusov, R. L., Koonin, E. V., Lipman, D. J., A Genomic Perspective on Protein Families *Science* **278**, 631-637 (1997).

[28] Taylor, W. R., Dynamic Sequence Databank Searching with Templates and Multiple Alignments. *J. Mol. Biol.*, **280**, 375-406 (1998).