Filtering erroneous protein annotation

*D. Wieser, E. Kretschmann and R. Apweiler**

Sequence Database Group, European Bioinformatics Institute, Cambridge, CB10 1SD, UK

ABSTRACT

Motivation: Automatically generated annotation on protein data of UniProt (Universal Protein Resource) is planned to be publicly available on the UniProt web pages in April 2004. It is expected that the data content of over 500 000 protein entries in the TrEMBL section will be enhanced by the output of an automated annotation pipeline. However, a part of the automatically added data will be erroneous, as are parts of the information coming from other sources. We present a post-processing system called Xanthippe that is based on a simple exclusion mechanism and a decision tree approach using the C4.5 data-mining algorithm.

Results: It is shown that Xanthippe detects and flags a large part of the annotation errors and considerably increases the reliability of both automatically generated data and annotation from other sources. As a cross-validation to Swiss-Prot shows, errors in protein descriptions, comments and keywords are successfully filtered out. Xanthippe is a contradictive application that can be combined seamlessly with predictive systems. It can be used either to improve the precision of automated annotation at a constant level of recall or increase the recall at a constant level of precision.

Availability: The application of the Xanthippe rules can be browsed at http://www.ebi.uniprot.org/

Contact: apweiler@ebi.ac.uk

1 INTRODUCTION

The protein databases Swiss-Prot, TrEMBL and the PIR are currently being unified into a single resource under the UniProt effort (Apweiler *et al.*, 2004). With the Swiss-Prot section of UniProt, users obtain a manually curated dataset of high qualitative value. Annotation is produced by a literature curation process and concerns mainly protein descriptions, i.e. names and synonyms, comments, keywords and sequence features. Large parts of the TrEMBL section, the non-curated remainder of UniProt, provide few, if any, of the above-mentioned value-adding annotation items. In this age of high-throughput sequencing, the manual curation process has not been able to cope with the avalanche of newly available sequence data, and as a consequence the proportion of well-annotated protein data is constantly shrinking. Since this situation is unsatisfactory, various applications to generate annotation automatically have been proposed as described in the literature (Prlic *et al.*, 2004; Fleischmann *et al.*, 1999 and others), and implemented in recent years.

One major aspect of the UniProt effort is to establish an automated annotation pipeline to provide users with predicted annotation, especially for otherwise little- or non-annotated database entries. The predictive annotation rule sets generated in the RuleBase (Biswas *et al.*, 2002) and Spearmint (Kretschmann *et al.*, 2001) projects are executed on a regular basis. The results of these approaches, which increase the data content of UniProt considerably, are expected to be presented to the general public on the project's Web pages from April 2004. They will be shown as prescriptive annotation on a separated layer and will suggest annotation without any modifications of the original data itself. It will be made obvious which annotation items were generated automatically and at which level of confidence they were produced.

However, cross-validation of predictive models against Swiss-Prot and various surveys of the produced data have shown that a part of automatically generated annotation is erroneous. The fact that both predictive systems applied in the automated annotation pipeline rely on protein families, domains and sequence signatures is one source of these errors. The InterPro database (Mulder *et al.*, 2003) provides these data by assigning a protein sequence to a particular domain or family based on the presence of a single signature hit. Whenever false positive hits are encountered, data mining applications have to deal with erroneous input data. It is a non-trivial task to render each and every annotation rule robust against this possibility. False positives appear over a wide range, with some hitting to related families or remotely similar biochemical properties, and some even occurring as entirely random events. Another source of errors lies in the bias between training and target sets, which are the Swiss-Prot and TrEMBL sections of UniProt, respectively. Some situations in the target set are not represented in the training set at all and can therefore not be resolved by mining algorithms using examples in the training set only. For instance, an annotation rule that was exported in Spearmint added the keyword 'Nuclear protein' to all entries in the TrEMBL section of UniProt having the InterPro domain IPR001005 ('Myb DNA-binding domain') and the SMART (Schultz *et al.*, 2000)

hit SM00717 ('SANT SWI3, ADA2, N-CoR and TFIIIB DNA-binding domains'). The keyword is annotated in all the 70 non-hypothetical Swiss-Prot proteins containing the InterPro domain and the SMART hit. In the target set however, protein Q819P5 ('Prespore specific transcriptional activator rsfA') fulfils the conditions of this annotation rule, despite its belonging to the kingdom of bacteria. There were no bacterial proteins in the training set and so the algorithm was not trained using a fully representative set of instances. Since there is no way of knowing what proteins are going to be present in a future TrEMBL database version, full representation of all circumstances in the training set can never be achieved.

Yet, a close analysis of the output of the annotation rule immediately leads to a straightforward method for filtering out this particular erroneous annotation. Bacteria do not possess nuclear proteins because of their lack of a nucleus. In all cases the 'Nuclear protein' keyword annotated on bacterial proteins is wrong, disregarding the origins of the annotation, which could be predictive systems, data imports or even human curation. This can be expressed as a simple exclusion rule, which if applied on the TrEMBL section of UniProt not only removes 66 wrong keyword predictions produced by automated annotation, but also spots the same error in some imports (e.g. in the bacterial protein Q93HH7).

In this paper, we present a system designed to mine automatically for exclusion rules to a far deeper level than in the above-mentioned obvious example, and to apply these rules to predicted, imported and literature curated annotations in UniProt database entries. The results of the application will be shown alongside the automated annotation part of the UniProt entry as prescriptive annotation. The project was named 'Xanthippe' after Socrates' renowned shrewish wife, due to the nature of the system to scrutinize the output of other systems and, if required, mark it as questionable.

2 SYSTEM AND METHODS

Both contradictors and predictors on protein data from UniProt use fundamental information entities present or precalculated for each protein. These are data such as the taxonomy of the organism, from which the protein was extracted, or the result of InterProScan (Zdobnov and Apweiler, 2001), which automatically classifies sequences into families and domains and detects hits to signature databases. This kind of information is considered to be core data and is available for each protein in UniProt.

It turned out that the presence and absence of annotation items is in many cases a function of the distribution of core data in the protein entry. The cases where annotation items are implied are mined by predictive systems, such as RuleBase and Spearmint. It is evident that there are also cases where annotation items are excluded by core data, but the predictors do not use this concept.

In the introductory example, the absence of 'Nuclear protein' keyword annotation in bacterial proteins was discussed, a fact that was deduced by using biological reasoning. In the following, two methods are presented which produce this and similar exclusion rules using a mere statistical data mining approach. They are eventually used as a system to avoid annotation errors in UniProt entries. Annotation items are marked as potentially wrong whenever the core data distribution in the entry suggests that the item should be absent.

Simple implication mechanism

The given example exploits the fact that organisms of the kingdom of bacteria do not possess any nuclear proteins. This simple implication can be detected automatically by examining the distributions of taxa (core data) and 'Nuclear protein' keywords (annotation) in the Swiss-Prot section of UniProt. Out of 138 920 entries, there are 56 967 bacterial and 8751 nuclear proteins. Assuming a normal distribution of these data items, an overlap would be expected, i.e. bacterial proteins with 'Nuclear protein' annotation. The expected value is 3589 instances, while the observed overlap is 0. In a database, where these data items are statistically distributed, this would be an extremely unlikely situation with a likelihood of $<1.4 \times 10^{-3388}$. This value is so small that not only can the assumption of a normal distribution be discarded but there is also a good indication that the two entities are mutually exclusive.

It is a simple task to design an algorithm that iterates through each taxon-keyword combination not present in Swiss-Prot and calculates a value for the probability of not observing an overlap. A threshold can be determined empirically, above which the combination is exported as an exclusion rule and can be applied on data in the TrEMBL section. At a threshold value of 1×10^{-10} around 4000 of such exclusions are generated.

Obviously, further mappings from core data can be used to contradict annotation for the corresponding proteins. Signature hits of the protein sequence and InterPro families or domains are particularly interesting and could be exploited to exclude annotation items. Yet, there are drawbacks to this approach. Proteins from a family or having a hit to a specific signature belong to their group according to specific properties. Unlike the set of bacterial proteins that covers a wide range of functions and hence contains a large number of distinct annotation items, the set of proteins belonging to a protein family by comparison contains a very limited range. In every protein family or group of proteins hitting a common sequence signature, most annotations are absent, and only a few are present. Literally millions of rules can be created and, in fact, the execution of these rules is far too inefficient in terms of application time.

Another drawback is that these exclusion rules will supposedly not detect a large proportion of the actual annotation errors. It is hardly likely that a high-quality prediction
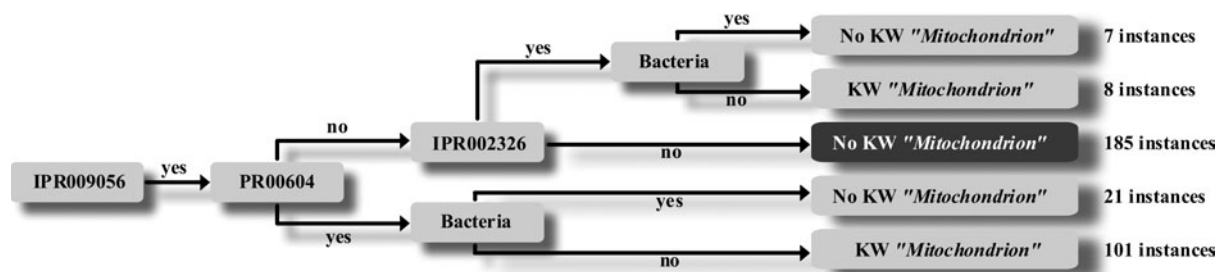
Fig. 1. Decision tree for the occurrence of 'Mitochondrion' keyword annotation in InterPro domain IPR009056.

mechanism or the literature curation process produces many annotations entirely non-specific to the protein families, to which a given target protein belongs. There is a better chance that annotation items, which are specific to the family of a target protein are affected by prediction errors. Unfortunately, such errors cannot be detected by using simple exclusions. This algorithm is designed to contradict only unspecific combinations, i.e. those which never occur in the given protein families. A better way of targeting them is to use a decision tree algorithm very similar to that employed in generating the Spearmint rule set.

Exclusion trees

While the mapping approach groups proteins globally into those having a core property and those not having it, exclusion trees are generated from a local and comparatively small set of training entities. The training sets are chosen to contain proteins that are reasonably similar to each other, for instance all Swiss-Prot entries belonging to a given InterPro family or domain. Because of the similarity between the proteins, the annotation in such groups is usually limited to a fairly small number of annotation items. Most errors will affect these items rather than those not occurring inside this group. To find a contradictor to prevent such errors, a decision tree for each annotation that occurs in a given training set is produced using the C4.5 algorithm (Quinlan, 1993). The leaves of the trees are examined to find the negative instances, i.e. those who do not have a particular annotation, and rules are derived, which describe the absence of the annotation. The set of all generated absence rules eventually serves as a contradictive system.

The following example illustrates the exclusion tree approach. Mining for the 'Mitochondrion' keyword in InterPro domain IPR009056 ('Cytochrome *c*') produces the decision tree shown in Figure 1. The tree is entirely generated on a statistical basis but it reflects some basic biological facts. In any case the prediction of the keyword is excluded from the kingdom of bacteria, who do not have mitochondria. More interesting for an exclusion rule generator are those proteins that neither hit to PRINTS (Attwood *et al.*, 2003) PR00604 ('Cytochrome *c*, class IA/ IB') nor belong to InterPro family IPR002326 ('Cytochrome *c*1'). Both families are found in

mitochondria, photosynthetic bacteria and other prokaryotes. The remainder of the InterPro domain IPR009056 ('Cytochrome *c*'), according to the decision tree, are not localized in the mitochondria.

A total of 185 instances in Swiss-Prot belong to InterPro IPR009056 and do not hit PRINTS PR00604 or InterPro IPR002326 (node with dark grey background in Fig. 1) and none of them has the 'Mitochondrion' annotation. For there to be a 'Mitochondrion' keyword predicted in this group, the protein would have to have at least one of the hits PR00604 or IPR002326; otherwise it will be contradicted by Xanthippe.

For the Swiss-Prot protein YIPP_DROME ('Yippee protein') from *Drosophila melanogaster*, the Spearmint system predicts a 'Mitochondrion' keyword. It uses a decision tree generated from training proteins in IPR000345 ('Cytochrome *c* heme-binding site'). This protein belongs to IPR009056 but hits neither PR00604 nor IPR002326, and hence, the keyword 'Mitochondrion' is not annotated in the original entry. Xanthippe exclusion trees detect this error made by Spearmint and mark it as possibly erroneous.

3 RESULTS

Since two inherently different sets of exclusion rules were generated, the results are presented separately to allow a thorough analysis of the individual systems.

Organism to keyword exclusions

This system is rather stable in its output, it being unlikely that in the future, organisms will be found that contradict the known fundamental properties of their taxonomical backgrounds. Since each data mining application on reasonably sized datasets produces false positive predictions, we chose to present the once exported exclusion rules to a biological expert who manually picked out those of biological value. One artefact that could be detected in this procedure was the apparent absence of ATP-binding proteins in a range of venomous snakes. The statistical approach produced a value far below the threshold and exported an exclusion between the taxonomy of these snakes and the 'ATP-binding' keyword. In reality this exclusion does not denote any exceptional metabolic properties of these animals, but more the high level of scientific
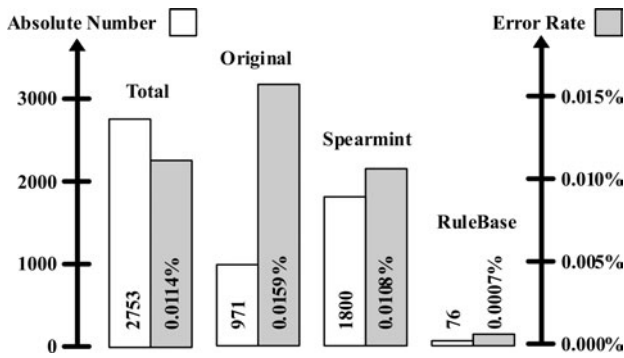
Fig. 2. Number of cases found in the TrEMBL section of UniProt, which were contradicted by Xanthippe exclusions from organism to keyword (white). Note, that the individual numbers do not add up, since overlaps occur. The grey part shows the error rate, i.e. how many contradictions where found compared with the amount of annotations provided by the individual sections.

interest in protein samples of their venom. None of these bind ATP, while the rest of the proteome is not represented in the Swiss-Prot section of UniProt.

In the following, the results of applying 700 curated rules are discussed. They were applied only when there was no example of that particular combination in Swiss-Prot. They were considered to be biological facts, so their confidence value was set to 100%. A cross-validation to Swiss-Prot is therefore unnecessary and only their performance in the TrEMBL section of UniProt and on automated annotation is given in Figure 2.

Spearmint was found to produce the largest amount of annotation errors in total numbers, while the original annotation on the entries, usually imports from EMBL (Kulikova *et al.*, 2004), had the highest error rate.

Exclusion trees

This method is highly sensitive to minor changes in the distribution of core data in the training set, and hence exclusion trees are exported with every release of a new version of Swiss-Prot. In general, the new exports differ from former results. They frequently query for related signature hits or on different levels in the taxonomy tree, but if applied they usually produce the same contradictions. Therefore, a curation step as in the first method is not feasible and the artefacts produced by this method have to be accepted and investigated.

Exclusion trees were generated for three inherently different annotation items: keywords, protein names and comments. Keywords consist of a controlled vocabulary of approximately 850 distinct words that are highly accessible to data mining algorithms. Protein names are less controlled and comments can be used for entirely free text annotation. Closer examination reveals that large parts of the latter annotation items in fact also consist of controlled vocabulary. They are kept consistent in the Swiss-Prot section of UniProt, occur in multiple
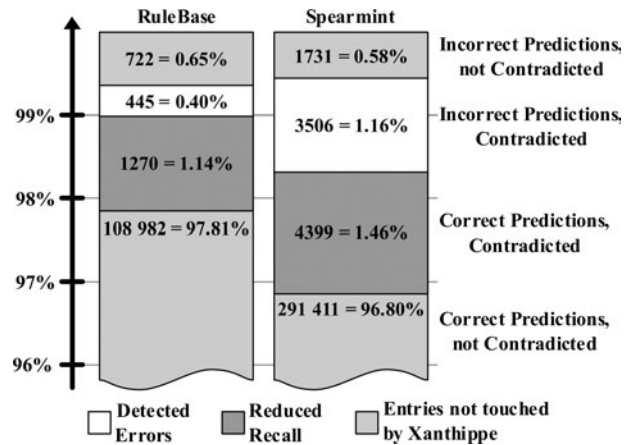
Fig. 3. Performance of Xanthippe exclusion trees on keyword predictions from RuleBase and Spearmint.

entries and can therefore be picked by algorithms working on a statistical basis.

The results are given individually for rules from Spearmint and RuleBase to show that the method performs well on predictive methods from entirely different backgrounds. RuleBase is an expert-curated annotation system, where the rules are created on the basis of biological reasons and only partly on statistical considerations. Spearmint ignores the biology component of the problem field, is founded on data mining algorithms only, and is related in its approach to the exclusion rule generator.

The cross-validations to the Swiss-Prot section of UniProt were sampled as follows. Whenever annotation rules from the individual sources were applied on a Swiss-Prot protein and a predicted annotation was not present in the entry itself, it was considered to be erroneous. Whenever this prediction was contradicted by the Xanthippe system, it was counted as detected (true positive); if it was not contradicted, it was counted as missed (false negative). There are cases where correct annotation was contradicted (false positive), but the largest part consists of correct annotations that were not contradicted (true negative).

Figure 3 shows the cross-validation of keyword predictions to Swiss-Prot. For Spearmint, approximately two-thirds of annotation errors are detected for the price of ~2% of wrongly contradicted annotations. If this validation is true for the target set, it suggests that if Spearmint predictions were applied to the TrEMBL section of UniProt physically mbox rather than by using the current prescriptive annotation system, the quality of keyword predictions could be increased from ~98.5 to ~99.5%. For RuleBase still ~40% of the annotation errors were found, improving the overall precision from 99.0 to 99.4%.

Figure 4 shows the Xanthippe performance on protein names and comments. For Spearmint, the protein name precision of the predictions could be increased from 96.9 to nearly
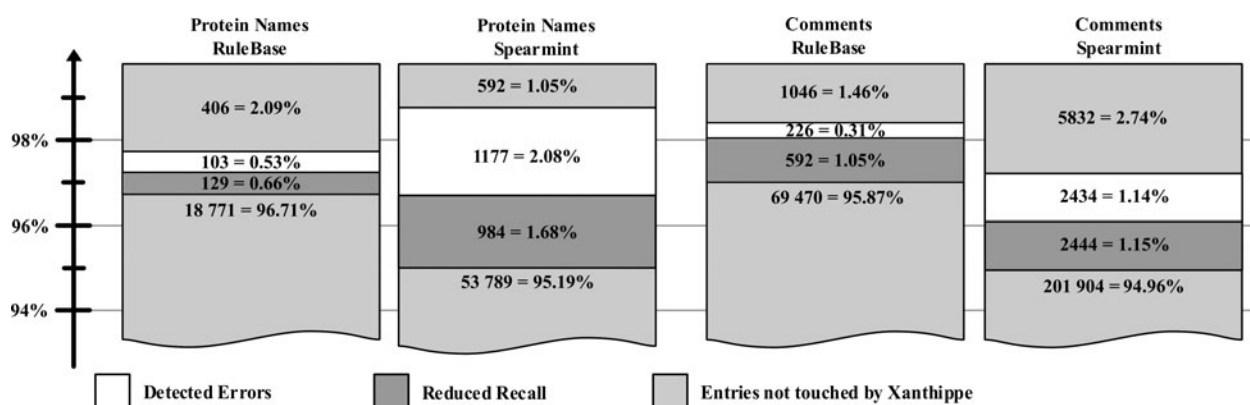
i345

Fig. 4. Performance of Xanthippe exclusion trees on comment and protein name predictions from RuleBase and Spearmint.

99%, all other items showed a poorer performance. After all, between 20 and 30% of the annotation errors could still be filtered on these data, but obviously these are the best targets for further improvements.

4   DISCUSSION

The presented system for filtering erroneous annotation from protein entries in UniProt, particularly the automatically produced data, proved to work on a sufficient level for keyword annotation. The filtering uses two distinct mechanisms, of which the simple mapping approach was approved to be taken into the automated annotation production pipeline. The exclusion tree approach turned out to be promising, but the performance of comment and protein name contradictions needs further improvement.

Improvements to the exclusion tree approach

Taking sub-string/super-string situations into account is expected to enhance the system considerably. In the InterPro family IPR000500 ('Connexins') for instance, all protein names of the Swiss-Prot members are annotated as 'Gap junction [extension] protein', where the extension can be 'alpha-1', 'beta-2', etc. The RuleBase system however predicts 'Gap junction protein' without the extension as protein name. Obviously, there can be no Xanthippe rule that ever contradicts this particular annotation, because there is no single instance in the training set where the extension is missing. Should there be a case where RuleBase annotates 'Gap junction protein' incorrectly, there is nothing in the current Xanthippe system that could prevent this from happening. Furthermore, the statistics given in the above diagrams are impaired by this effect. In the cross-validation, 'Gap junction protein' predictions on actual 'Gap junction [extension] protein' annotations are still considered as false positives. Mostly, sub-string predictions are generalizations of the actual annotation and should be calculated as true positives. The Xanthippe system needs to be extended to cover sub-strings

of protein names and comments, and hence these cases need to be included in the training sets.

Predictors versus contradictors

The RuleBase system has been used since 1999 to produce automated annotation. The performance in terms of precision is highly appreciated and the system is supported by a team of experts. Spearmint is intended to supplement RuleBase and to increase its coverage without compromising its level of precision.

As an automatic data mining system Spearmint has the advantage of producing much more annotation than RuleBase. The latter however is small enough to be reviewed constantly by scientists, and hence the confidence in the data it produces is high. The output quality of Spearmint can be adjusted by not applying all the exported rules, but only those with high statistical support. Taking RuleBase as a benchmarking system, Spearmint is set to produce the same quality level as RuleBase. The Spearmint application, followed by a Xanthippe postprocessor, can be adjusted to produce even more predictions at a lower level of precision. If a large portion of the errors is filtered out, the same overall annotation quality can be produced as in a more restrictive Spearmint export without the Xanthippe post-processing step.

Running Spearmint at 98.5% quality reproduces 33% of keyword annotation in Swiss-Prot (Kretschmann *et al.*, 2001), while a 95% quality level yields 58% recall. Provided that Xanthippe still detects two-thirds of the errors produced by a Spearmint exported at 95%, the actual precision is expected to be at around the desired 98.5%. Additionally, the proportion of detected erroneous annotation is expected rather to increase for a rule set produced by a comparatively low precision system. This means that with using Xanthippe the recall of keywords can be nearly doubled without compromising the quality of the prediction.

Feedback loops

If RuleBase or Spearmint predicts an annotation item on a Swiss-Prot entry, which is not contradicted by Xanthippe

but is missing in the entry, a further investigation should be undertaken. In some cases the item might be missing in the entry and can be added, which would improve data consistency in Swiss-Prot.

Since garbage in–garbage out effects are responsible for many annotation errors, the signatures leading to the most obvious ones will be reported to InterPro. If required, such hits can be set to false positive status in this database.

5 CONCLUSION

This work shows that predictive models can be enhanced by including an additional contradictive level. The approach works sufficiently well for a mining environment on protein data. The authors suspect that other data mining applications could benefit from similar systems.

ACKNOWLEDGEMENTS

REFERENCES

Apweiler,R., Bairoch,A., Wu,C.H., Barker,W.C., Boeckmann,B., Ferro,S., Gasteiger,E., Huang,H., Lopez,R., Magrane,M. *et al.* (2004) UniProt: the Universal Protein Knowledgebase. *Nucleic Acids Res.*, **32**, D115–D119.

Attwood,T.K., Bradley,P., Flower,D.R., Gaulton,A., Maudling,N., Mitchell,A.L., Moulton,G., Nordle,A., Paine,K., Taylor,P., Uddin,A. and Zygouri,C. (2003) PRINTS and its automatic supplement, prePRINTS. *Nucleic Acids Res.*, **31**, 400–402.

Biswas,M., O'Rourke,J.F., Camon,E., Fraser,G., Kanapin,A., Karavidopoulou,Y., Kersey,P., Kriventseva,E., Mittard,V., Mulder,N. *et al.* (2002) Applications of InterPro in protein annotation and genome analysis. *Brief. Bioinform.*, **3**, 285–295.

Fleischmann,W., Möller,S., Gateau,A. and Apweiler,R. (1999) A novel method for automatic functional annotation of proteins. *Bioinformatics*, **15**, 228–233.

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

Kulikova,T., Aldebert,P., Althorpe,N., Baker,W., Bates,K., Browne,P., Van Den Broek,A., Cochrane,G., Duggan,K., Eberhardt,R. *et al.* (2004) The EMBL Nucleotide Sequence Database. *Nucleic Acids Res.*, **32**, D27–D30.

Mulder,N.J., Apweiler,R., Attwood,T.K., Bairoch,A., Barrell,D., Bateman,A., Binns,D., Biswas,M., Bradley,P., Bork,P. *et al.* (2003) The InterPro Database, 2003 brings increased coverage and new features. *Nucleic Acids Res.*, **31**, 315–318.

Prlic,A., Domingues,F.S., Lackner,P. and Sippl,M.J. (2004) WILMA—automated annotation of protein sequences. *Bioinformatics*, **20**, 127–128.

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

Schultz,J., Copley,R.R., Doerks,T., Ponting,C.P. and Bork,P. (2000) SMART: a web-based tool for the study of genetically mobile domains. *Nucleic Acids Res.*, **28**, 231–234.

Zdobnov,E.M. and Apweiler,R. (2001) InterProScan—an integration platform for the signature-recognition methods in InterPro. *Bioinformatics*, **17**, 847–848.