Selection of Data Sets of Motifs as Attributes in the Process of Automating the Annotation of Proteins' Keywords

Ana L.C. Bazzan* and Cassia T. dos Santos

Instituto de Informática, Universidade Federal do Rio Grande do Sul,
Caixa Postal 15064, 91.501-970, Porto Alegre, RS, Brazil
{bazzan, ctsantos}@inf.ufrgs.br

1 Introduction and Related Work

Automatic annotation tools are becoming popular since the biologists and curators of databases cannot cope with the volume of sequences to be annotated manually. One way to automate the annotation is to use techniques of symbolic machine learning to derive rules to guide this annotation. However, the training instances tend to have too many attributes, turning the machine learning process difficult and time consuming.

The aim of this paper is to evaluate the information provided by those attributes, which can come from different data sets, regarding a simple task: classifying proteins according to a given set of keywords. Despite its simplicity, the task is very relevant because the Keyword field is an important one in the SWISS-PROT database and gives several hints to experts regarding proteins function and structure. Instead of using thousands of attributes during the machine learning process, we study which set of these attributes can potentially contribute more to the annotation process. Once those rules are generated, they are used to fill the Keyword field in the TrEMBL database (a computer-annotated supplement of SWISS-PROT).

The idea of automating the annotation is not new. Machine learning tecnhiques have been widely used in automated annotation process. An approach based on these techniques to generate rules based on already annotated keywords of the SWISS-PROT database is described by [3]. Such rules can then be applied to unannotated protein sequences in TrEMBL.

In [1] similar methods were employed to automate the annotation of Keyword for proteins appearing in the genome of organisms of the *Mycoplasmataceae* family. However, as said, one issue with this approach is that it still uses too many attributes (all motifs from InterPro and PROSITE cross-referenced in SWISS-PROT). We believe that the time consumed in the training task can be reduced if the correct set of attributes is used.

* Authors partially supported by CNPq; the project is partially supported by CNPq and by FAPERGS.

J.C. Setubal and S. Verjovski-Almeida (Eds.): BSB 2005, LNBI 3594, pp. 230–233, 2005.

2 Data and Methods

Here we use data about proteins from the model organism *Arabidopsis thaliana*, which is available in SWISS-PROT, to feed the Layer II of ATUCG, our agent-based environment for annotation [2]. SWISS-PROT[1] provides a high level of annotation of each protein, also including extensive cross-references to other databases of motifs, patterns, and profiles. We use some of these cross-references as attributes in the machine learning process. Specifically, we use the following. PROSITE[2] characterizes biologically significant sites in proteins. Pfam[3] is a database of alignments and HMMs covering many common protein domains. PRINTS[4] is a compendium of protein fingerprints. ProDom[5] families are built by an automated process based on a recursive use of PSI-BLAST. Finally, InterPro[6] uses a collection of profiles from PRINTS, Prosite, ProDom, Pfam and SWISS-PROT, which creates a unique, non-redundant characterization of a given protein family, domain or functional site.

The data used comes from a local version of the SWISS-PROT database (status of May, 2004), in which 2817 proteins relating to *A. thaliana* were found. Many keywords appeared in the data but we are focusing on those whose number of instances is higher than 100. The number of keywords satisfying this criterion is 27 (those that appear in Table 1). Since the aim here is to compare data sets of motifs we use all motifs which are cross-referenced in SWISS-PROT as attributes. The number of attributes, by data set, is: 1316 (Intepro), 907 (Pfam), 220 (Prodom), 589 (Prosite), 246 (Prints), thus 3278 in total. Also, we have imposed a constraint on the quality of the rules generated by C4.5: each rule must cover a minimum number of 25 instances, a number that is approximately 1% of the number of training instances. The quality of each rule generated by C4.5 was evaluated via 5-fold cross-validation (CV).

3 Results and Discussion

In Table 1, the first column is a list of the keywords which met the above mentioned criteria. The second column gives the global error. The third and fourth blocks of columns relate to the statistics for the positive and the negative classes respectively. In these two blocks, averages (due to the n-fold CV) of the number of instances, the absolute error, and the percentage of error are shown. Also, for the positive class only, the table shows confidence as defined in [3].

Due to lack of space, we omit the other tables, showing in Table 2 only the equivalent of the last line of Table 1 (average over all keywords). When the

[1] http://www.expasy.ch/sprot/
[2] http://www.expasy.ch/prosite
[3] http://www.sanger.ac.uk/Pfam/
[4] http://bioinf.mcc.ac.uk/dbbrowser/PRINTS/PRINTS.html
[5] http://protein.toulouse.inra.fr/prodom.html
[6] http://www.ebi.ac.uk/interpro/

Table 1. Evaluation Test (5-fold CV) - Attributes Used: Interpro

	Global	Class (Keyword)			Non-Class
Keyword	Error (%)	Instances	Error (%)	Conf.	Error (%)
ATP-binding	21.60 (3.80)	53.6	21.20 (39.55)	0.87	0.40 (0.08)
Alternative-splicing	27.20 (4.80)	27.2	27.20 (100.00)	0.00	0.00 (0.00)
Calcium	8.20 (1.40)	22.2	7.80 (35.14)	0.75	0.40 (0.07)
Cell-wall	8.80 (1.60)	24.2	7.80 (32.23)	0.74	1.00 (0.19)
Chloroplast	71.00 (12.60)	71	71.00 (100.00)	0.00	0.00 (0.00)
Coiled-coil	23.60 (4.20)	33	21.00 (63.64)	0.57	2.60 (0.49)
DNA-binding	33.00 (5.80)	47.4	33.00 (69.62)	0.79	0.00 (0.00)
Glycoprotein	31.80 (5.70)	49.2	29.80 (60.57)	0.72	2.00 (0.39)
Heme	4.00 (0.70)	32.6	3.80 (11.66)	0.87	0.20 (0.04)
Hydrolase	36.80 (6.50)	51.8	36.60 (70.66)	0.78	0.20 (0.04)
Iron	13.00 (2.30)	27.4	12.80 (46.72)	0.77	0.20 (0.04)
Metal-binding	23.20 (4.10)	38.6	23.00 (59.59)	0.78	0.20 (0.04)
Mitochondrion	44.20 (7.80)	44.2	44.20 (100.00)	0.00	0.00 (0.00)
Multigene-family	152.60 (27.10)	70	0.00 (0.00)	0.86	4.60 (1.33)
Nuclear-protein	55.40 (9.80)	76.8	55.40 (72.14)	0.85	0.00 (0.00)
Oxidoreductase	34.80 (6.20)	63.4	34.60 (54.57)	0.87	0.20 (0.04)
Phosphorylation	15.20 (2.70)	25.8	11.20 (43.41)	0.56	4.00 (0.74)
Plant-defense	12.00 (2.20)	23.4	12.00 (51.28)	0.75	0.00 (0.00)
Protein-transport	21.40 (3.80)	23.2	21.40 (92.24)	0.32	0.00 (0.00)
Repeat	42.60 (7.50)	62.4	42.60 (68.27)	0.84	0.00 (0.00)
Ribosomal-protein	34.00 (6.00)	34	34.00 (100.00)	0.00	0.00 (0.00)
Signal	60.20 (10.70)	98.8	59.20 (59.92)	0.87	1.00 (0.22)
Transcription-regulation	26.00 (4.60)	47.4	26.00 (54.85)	0.85	0.00 (0.00)
Transferase	43.20 (7.70)	57.4	43.00 (74.91)	0.77	0.20 (0.04)
Transit-peptide	69.00 (12.30)	69	69.00 (100.00)	0.00	0.00 (0.00)
Transmembrane	79.60 (14.10)	111.8	78.40 (70.13)	0.84	1.20 (0.27)
Transport	40.40 (7.20)	48.2	40.40 (83.82)	0.67	0.00 (0.00)
Average	38.25 (6.79)	49.41	32.09 (63.51)	0.62	0.68 (0.15)

classification is performed with attributes only from single databases in Table 2, in most cases the error in the non-class is low. However, looking at error rates regarding the positive class only (fourth column), some are unacceptable (e.g. 95.07% for ProDom). Similar conclusion can be drawn for confidence. If we consider attributes only from the InterPro database, we see that the error rate in the positive class is lower than it was the case when only ProDom was used. This is valid for all keywords (though not shown here).

For the other data sets, the trend is that global error is low (e.g. 7.75% for PRINTS) but the error rate for the positive class is high. Better confidences and error rates are achieved when using the following databases: InterPro, Pfam, and also for the combinations: InterPro+PROSITE, and InterPro+PROSITE+Pfam). However, in these last cases, the combination brought no increase: using attributes from InterPro alone is as good as using attributes from InterPro plus other data sets.

Table 2. Evaluation Test (5-fold CV) – Error, number of instances and confidence, for each data set of attributtes (average over all keywords)

	Global	Class (Keyword)			Non-Class
Database	Error (%)	Instances	Error (%)	Conf.	Error (%)
All	37.98 (6.74)	49.50	31.94 (63.31)	0.62	0.65 (0.14)
Interpro+Prosite+Pfam	37.98 (6.74)	49.51	31.90 (63.19)	0.62	0.70 (0.15)
Interpro+Prosite	37.96 (6.73)	49.51	31.88 (63.16)	0.62	0.70 (0.15)
Interpro	38.25 (6.79)	49.41	32.09 (63.51)	0.62	0.68 (0.15)
Pfam	39.52 (7.01)	49.21	33.13 (65.38)	0.60	0.68 (0.15)
Prosite	40.35 (7.15)	49.32	34.32 (68.92)	0.57	0.44 (0.09)
Prints	43.70 (7.75)	48.64	37.13 (73.55)	0.47	0.31 (0.06)
Prodom	52.54 (9.32)	50.36	47.77 (95.07)	0.20	0.23 (0.04)

Finally, a note on the still high level of error rate. This is due to two main factors: low level of annotation of Keyword in SWISS-PROT and the unbalance of the two classes. This issues were investigated somewhere else and are not the focus of the present paper, which aims at comparing the data sets.

4 Conclusions

Using all available data regarding motifs as attributes is prohibitive for symbolic machine learning methods. This paper discusses the use of several data sets in order to evaluate which one(s) is/are more valuable regarding the task of producing rules for annotation of the field Keyword in TrEMBL.

One sees that some data sets of attributes perform similarly. In particular, using all attributes (i.e. from all databases together) does not perform better than using only InterPro or only Pfam. Combinations of attributes (e.g. PROSITE+InterPro or PROSITE+InterPro+Pfam) do not perform much better than each of these data sets alone. ProDom or PRINTS should not be used alone as data set in the automated techniques, at least at this time when the data set is small. Since each of these databases has its particularities, the expert in the domain of annotation should decide which one to use. In the absence of this information, InterPro is a safe choice since it is based on the others.

References

1. A. L. C. Bazzan, S. C. da Silva, P. M. Engel, and L. F. Schroeder. Automatic annotation of keywords for proteins related to *Mycoplasmataceae* using machine learning techniques. *Bioinformatics*, 18(S2):S1–S9, October 2002.
2. A. L. C. Bazzan, R. Duarte, A. N. Pitinga, S. L. F., S. C. Silva, and F. A. Souto. ATUCG–an agent-based environment for automatic annotation of genomes. *International Journal of Cooperative Information Systems*, 12(2):241–273, June 2003.
3. E. Kretschmann, W. Fleischmann, and R. Apweiler. Automatic rule generation for protein annotation with the C4.5 data mining algorithm applied on SWISS-PROT. *Bioinformatics*, 17:920–926, 2001.