PROTEIN FAMILY CLASSIFICATION WITH DISCRIMINANT FUNCTION ANALYSIS

Etsuko N. Moriyama and Junhyong Kim[□]

1. INTRODUCTION

Rapid progress in multiple genome projects continues to feed databases in the world a large volume of sequence data. In this 'post-genomic' era, more efficient and reliable sequence annotation, especially functional annotation of protein sequences, is crucial. Although experimental confirmation is ultimately required, computational annotation of protein sequences has been routinely done, and it is incorporated into major protein databases (*e.g.*, SWISS-PROT: http://www.expasy.org/sprot/, PIR-PSD: http://pir.georgetown.edu/pirwww/search/textpsd.shtml). Due to a rapidly growing number of new sequences, increasingly more database entries contain only computational annotations.

In this paper, we first discuss the disadvantage commonly found in various existing protein classification methods. Next we introduce a set of new methods that can classify protein family sharing very weak similarity. Finally, we describe an algorithm that combines strengths from various protein classification methods to obtain an optimum power for protein classifications.

1.1. Protein Classification Methods

In order to improve the power of computational annotations, various methods have been developed. Computational annotation (or classification) methods rely on finding similarity between a query (new protein) sequence and protein sequences in databases with known (preferably experimentally confirmed) functions. The most popularly used method is the Basic Local Alignment Search Tool (BLAST) by Altschule *et al.* (1990). It searches databases for sequences with local similarity to the query. When more distant similarity is sought, pattern or profile, rather than the sequence itself, is used for the database search. Table 1 lists some methods frequently used for protein annotation and

[□] Etsuko N. Moriyama, School of Biological Sciences and Plant Science Initiative, University of Nebraska, Lincoln, Nebraska, 68588-0660. Junhyong Kim, Department of Biology, University of Pennsylvania, Philadelphia, Pennsylvania, 19104-6018.

protein family classification.

Table 1. Protein classification tools

Tool	Description	Reference
BLAST	Local sequence similarity search tools (blastn, blastp, *etc.*)	Altschule *et al.* (1990)
PRINTS/SPRINT	Protein fingerprint database (searched by FingerPRINTScan)	Attwood *et al.* (2002)
PROSITE	Database for biologically significant sites, patterns and profiles	Falquet *et al*. (2002)
Pfam	Multiple alignment and profile HMM database (searched by HMMER)	Bateman *et al*. (2002)
PSI-BLAST	Position specific iterative BLAST using position specific scoring matrix	Altschul *et al*. (1997)
SMART	Domain architecture research tool (profile HMM database)	Letunic *et al*. (2002)

Except BLAST (including PSI-BLAST), all of the search tools listed above have their own pattern, profile, or motif database. These patterns/profiles are generated from alignments of known protein sequences. Since functionary more important regions (*e.g.*, catalytic domains, binding-domains) are considered to be under stronger selective constraints, multiple alignments from proteins with known functions are expected to contain conserved regions related to those functions. When distantly related sequences are compared, only functionally crucial sites, rather than a large region, could be conserved. Furthermore, some amino acids could be substituted with others (usually with other bio-chemically similar amino acids) as long as the protein function is maintained. Pattern and profile search methods allow such flexibility and more sensitive to weakly conserved sequences than simple similarity search methods. Even when regular BLAST search fails to identify any significantly similar sequence to the query from the database, frequently pattern/profile search methods can detect a signature pattern/profile related to a known function.

1.2. Pros and Cons for the Current Protein Classification Tools

Due to the differences in their underlying techniques and also in their focuses (*e.g.*, family coverage), each method (and database) has different strengths and weaknesses. In order to take maximum advantage from these various information sources, usually it is necessary to conduct multiple pattern/profile searches. Integrated databases, *e.g.*, InterPro (http://www.ebi.ac.uk/interpro/) and MetaFam (http://metafam.ahc.umn.edu/), were developed to facilitate such tedious procedures.

One of the problems inherited in all of these pattern/profile search methods and databases is that their patterns are in general derived from relatively short regions. It is particularly the case in the PROSITE patterns. PROSITE patterns are expressed in regular expressions as shown in Figure 1. If the query is only a partial sequence (*e.g.*, EST), and if it does not contain the region where the pattern was derived, this method fails to identify the query correctly.

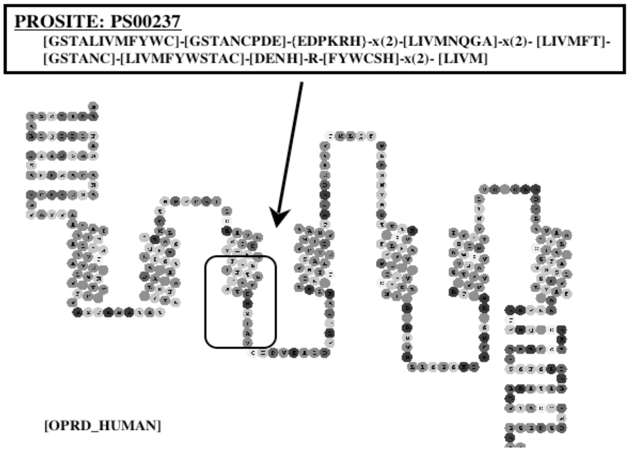

PROSITE: PS00237
[GSTALIVMFYWC]-[GSTANCPDE]-(EDPKRH)-x(2)-[LIVMNQGA]-x(2)- [LIVMFT]-[GSTANC]-[LIVMFYWSTAC]-[DENH]-R-[FYWCSH]-x(2)- [LIVM]
[OPRD_HUMAN]

Figure 1. An example entry from the PROSITE pattern database. This is one of the pattern entries for G-protein coupled receptors (PS00237). A G-protein couple receptor sequence (a human opioid receptor entry from SWISS-PROT: OPRD_HUMAN) is shown under the PROSITE pattern entry. Each dot represents one amino acid. Seven cylinders in the middle indicate predicted seven transmembrane regions. The circled area on the sequence corresponds to the PROSITE regular expression pattern entry spanning 17 amino acids.

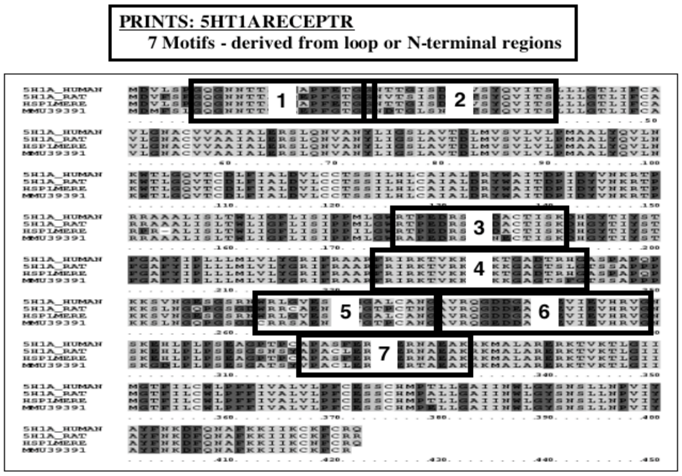
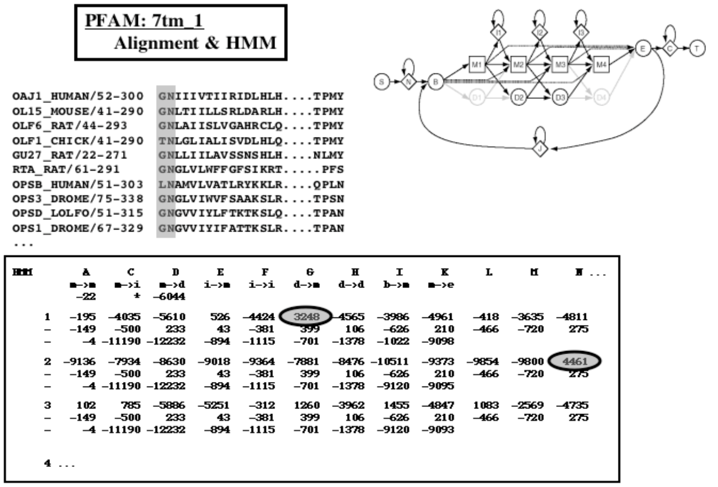


Figure 2. An example protein fingerprint from the PRINTS database. This is "5-hydroxytryptamine 1A receptor signature" (5HT1ARECEPTR). This PRINTS entry includes seven motifs or "fingerprints." The alignment below (including four 5-hydroxytryptamine 1A receptor sequences) shows the locations of these seven fingerprints (the boxes 1 – 7).

PRINTS uses also very short conserved motifs. But it tries to overcome this problem by identifying multiple motifs covering a larger region than a single motif. Figure 2 in the previous page shows one example PRINTS entry that includes seven fingerprint motifs.

Profiles (used in PROSITE/Profile and PSI-BLAST) express the flexibility in amino acid substitutions for each position in a series of scoring matrices, "Position Specific Scoring Matrix" (PSSM). The profile hidden Markov model (profile HMM) is a probabilistic model of sequences and used in Pfam and SMART. Profiles and profile HMMs cover the entire region of alignments, much longer than regions covered by PROSITE/pattern or PRINTS. Figure 3 below shows an example profile HMM from a Pfam entry.


PFAM: 7tm_1
Alignment & HMM

```
OAJ1_HUMAN/52-300   GNIIIVTIIRIDLHLH....TPMY
OL15_MOUSE/41-290   GNLTIILLSRLDARLH....TPMY
OLF6_RAT/44-293     GNLAIISLVGAHRCLQ....TPMY
OLF1_CHICK/41-290   TNLGLIALISVDLHLQ....TPMY
GU27_RAT/22-271     GNLLIILAVSSNSHLH....NLMY
RTA_RAT/61-291      GNGLVLWFFGFSIKRT.....PFS
OPSB_HUMAN/51-303   LNAMVLVATLRYKKLR....QPLN
OPS3_DROME/75-338   GNGLVIWVFSAAKSLR....TPSN
OPSD_LOLFO/51-315   GNGVVIYLFTKTKSLQ....TPAN
OPS1_DROME/67-329   GNGVVIYIFATTKSLR....TPAN
...
```

```
HMM        A      C      D      E      F      G      H      I      K      L      M      N ...
        m->m   m->i   m->d   i->m   i->i   d->m   d->d   b->m   m->e
         -22      *  -6044

    1   -195  -4035  -5610    526  -4424   3248  -4565  -3986  -4961   -418  -3635  -4811
    -   -149   -500    233     43   -381    399    106   -626    210   -466   -720    275
    -     -4 -11190 -12232   -894  -1115   -701  -1378  -1022  -9098

    2  -9136  -7934  -8630  -9018  -9364  -7881  -8476 -10511  -9373  -9854  -9800   4461
    -   -149   -500    233     43   -381    399    106   -626    210   -466   -720    275
    -     -4 -11190 -12232   -894  -1115   -701  -1378  -9120  -9095

    3    102    785  -5886  -5251   -312   1260  -3962   1455  -4847   1083  -2569  -4735
    -   -149   -500    233     43   -381    399    106   -626    210   -466   -720    275
    -     -4 -11190 -12232   -894  -1115   -701  -1378  -9120  -9093

    4 ...
```

Figure 3. An example entry from Pfam. A part of a G-protein coupled receptor entry (7tm_1) is shown in the box. Two circled numbers correspond to the "emission probability" for a glycine at the first amino acid position and an alanine at the second position, respectively. Note that the amino acids "G" and "N" are the majority at the first and second positions in the alignment above, respectively, and the "emission probabilities" for these two amino acids are the largest at each site. The diagram above is the transition structure of the HMM model used in HMMER (the HMM program package used in Pfam).

Another inherent problem in these methods is that they rely on multiple alignments for generating the patterns and profiles. However, making multiple alignments themselves becomes problematic when extremely distant similarity is on the issue. Furthermore, diagnostic patterns and profiles cannot be easily identified from such multiple alignment including extremely diverged sequences.

Yet another problem shared by existing classification methods is that the patterns, motifs, and profiles need to be identified from already known protein sequences. Since subsequently found proteins are classified based on these patterns/profiles, possible initial sampling bias will be reinforced.

1.3. G-protein Coupled Receptor Super Family

One good example representing the case of extreme diversity is the G-protein coupled receptor (GPCR) super family. Many medically and pharmacologically important proteins are included in this family: *e.g.*, acetylcholine receptors, dopamine receptors, and opioid receptors. Therefore, classifying this protein family and finding new members of this family is one of the most important topics in medical genomics.

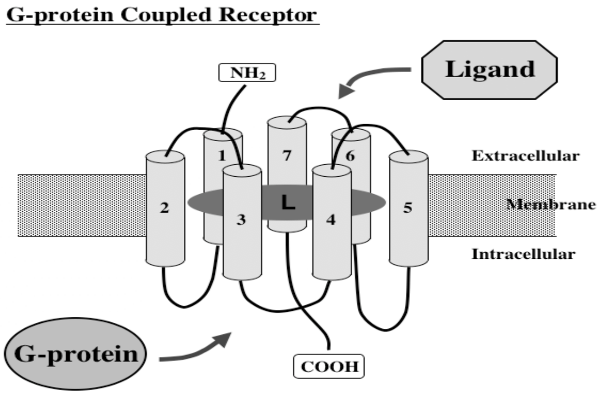


Figure 4. A model of G-protein coupled receptors. The seven transmembrane regions are shown in cylinders with numbers 1 – 7.

The GPCR protein family is one of the most diverse protein families. The family is classified into five major classes (A-E) as well as other minor classes and putative and "orphan" groups. The members of this family share one structural feature, seven-transmembrane regions as shown in Figure 4. Beyond this structural similarity, the members, especially those in different classes share very low sequence similarity. The seven transmembrane regions contribute to the low sequence similarity because many of the hydrophobic amino acids within the region are interchangeable as long as they are hydrophobic and do not disrupt the structural conformation. On the other hand, the loop regions between the transmembrane regions can be varied in length. Therefore, transmembrane and loop regions contribute to the low sequence similarity in different ways. The low sequence similarity and heterogeneity created by repeated transmembrane and loop regions creates the most difficult situation in generating multiple alignments, and no reliable multiple alignment can be generated from the entire super family. The GPCR protein family presents one of the most challenging properties for protein classification methods. There is no single pattern or profile representing the entire GPCR family. And even if the query is predicted to belong to the GPCR super family, extreme divergence among families sometimes prevents further classification. In such cases, the query sequences are called "orphan" GPCRs. For example, the current GPCRDB ("Information system for G protein-coupled receptors (GPCRs)": http://www.gpcr.org/7tm/) contains about 4,650 GPCR entries, and more than 300 entries are designated as "orphans" or "putative/unclassified" GPCRs.

At varied degrees, such situations are shared with many other transmembrane proteins. We can expect that when we develop methods that can successfully classify this particular protein family, such methods can be applied easily for many other protein families. We therefore use the GPCR protein family in our study to evaluate performance of various classification methods.

2. DISCRIMINANT ANALYSIS

In Kim *et al.* (2000), we described a new method of protein classification that relies on neither multiple alignments nor pattern/profile database search. The new method uses a set of variables extracted from each protein sequence, and classifies them by using a "nonparametric" linear discriminant analysis. It is a linear discriminant analysis optimized with nonparametric "runs" criterion, instead of relying on parametric equations commonly used. This was because we wanted to avoid assuming any unreasonable statistical distribution. In this paper, we include other parametric and nonparametric discriminant analysis methods and compared their performance for the GPCR family classification.

2.1. Input Variables

Instead of using multiple alignments, a set of variables extracted from each protein sequence is used in discriminant analyses. We used the same set of variables described in Kim *et al.* (2000). They include "amino acid index" and three periodicity statistics based on hydrophobicity and polarity.

In order to obtain "amino acid indices" from protein sequences, first a linear discriminant analysis is done based on 19 amino acid frequencies. The linear discriminant score (we call it "amino acid index") obtained from each protein sequence can be used as a single variable representing 19 amino acid frequencies. Note that using "amino acid index" instead of each amino acid frequency separately reduces the dimensionality from 19 to 1.

Other three variables are:

i) Log of the average periodicity of the GES scale,
ii) Log of the average periodicity of the polarity scale, and
iii) Variance of first derivative of the polarity scale.

Distributions of the GES hydropathy index (Engelman *et al.*, 1986) and polarity along each protein sequence are examined using sliding window analysis (window size = 16 amino acids). And the "average periodicity" is calculated by counting how many times the property crosses over a neutral value (-0.5 for the GES and 8.325 for the polarity) and normalized by protein lengths.

Figure 5 in the next page shows how these four input variables discriminate GPCRs from non-GPCRs. Note that although the figure shows only two-dimensional variable spaces, good discriminations between GPCR proteins (shown with "G" in the figure) and non-GPRCR proteins (shown with "R" in the figure) were observed from any combination of the four variables.

2.2. Datasets

A training dataset containing 750 of known GPCR sequences (randomly sampled from GPCRDB) and 1,000 of non-GPCR sequences (randomly sampled from SWISS-PROT) were prepared. A smaller dataset including 100 each of GPCR and non-GPCR sequences were also prepared independently as a test dataset.

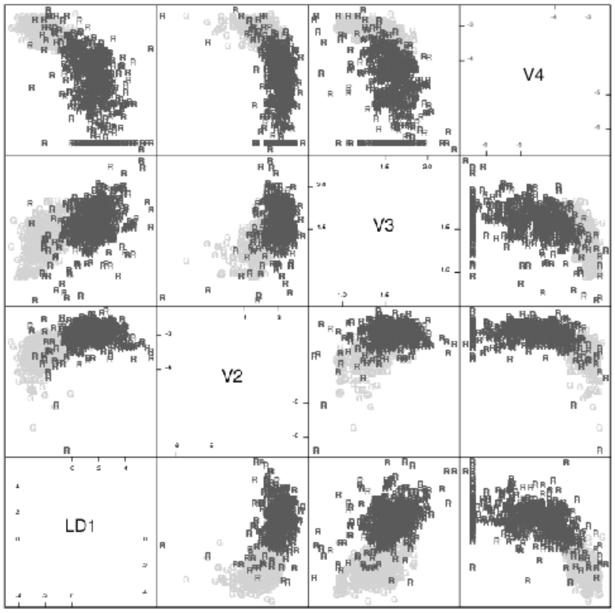


Figure 5. A multi-plot for the four input variables used in discriminant analyses. LD1: "amino acid index", V2, V3, and V4: hydrophobicity and polarity periodicity statistics. 1,750 proteins in the training dataset (described below) are plotted. "G" in grey color: GPCRs, "R" in black color: non-GPCR random proteins.

2.3. Discriminant Analysis Methods

In addition to the "nonparametric" linear discriminant analysis (nonparametric LDA) method described previously (Kim *et al*., 2000), we included four other parametric and nonparametric discriminant analysis methods:

i) Linear discriminant analysis (LDA),
ii) Quadratic discriminant analysis (QDA),
iii) Logistic discriminant analysis (LOG), and
iv) K-nearest neighbor method (KNN).

S-Plus statistical package with the MASS library (Venables and Ripley, 2002) was used except for the nonparametric LDA.

2.4. Performance Comparisons for the GPCR Protein Family Classification

Table 2 lists the results of GPCR classification by various discriminant analyses compared with other protein classification methods. Discriminant functions were trained on the training dataset described above, and their classification performance was tested on both of the training and test datasets. The cross-validation ("leave-one-out" test) were performed only for the four parametric and nonparametric discriminant analyses.

Table 2. Performance comparisons for the GPCR classification

Method	Against training dataset		Against test dataset		Cross-validation	
	% True +	% False +	% True +	% False +	% True +	% False +
LDA	98.7	3.3	100	1	98.7	3.6
QDA	98.5	3.0	100	0	98.5	3.0
LOG	98.0	2.8	100	0	97.7	2.9
KNN (k = 10)	98.7	3.2	99	0	98.3	3.4
Nonparametric LDA	98.1	3.6	99	0	-	-
PROSITE/pattern	93.5	0.1	84	0	-	-
PROSITE/profile	98.8	0	94	0	-	-
Pfam	98.4	0.3	94	0	-	-
PRINTS	99.2	0.3	98	0	-	-

NOTE: "%True +": percent true positives, "% False +": percent false positives. For KNN analysis, k was varied between 5 and 20, but the performance with k=5, 10, and 15 was similar. See Kim *et al*. (2002) for the PROSITE, Pfam, and PRINTS entries used in the analyses.

All of the four parametric and non-parametric methods performed similar to or better than previously described methods (PROSITE, Pfam, PRINTS, and nonparametric LDA) with % true positives higher than 98 %. Surprisingly both non-parametric methods, KNN and nonparametric LDA, did not perform particularly better than parametric discriminant analysis (DA) methods, although we cannot guarantee any assumption underlying parametric methods (*e.g.*, normal distribution, consistent covariance matrices). The false positive rates among DA methods were also very similar to each other, but about 10 times higher than other methods (PROSITE, Pfam, and PRINTS). Classification of the independently prepared test dataset and the results of cross-validation were consistent with the classification results on the training dataset itself. As described before, classification by PROSITE/pattern search showed the lowest performance. It clearly indicates the limitation of searching short patterns.

On the other hand, all of the DA methods outperformed other methods when tested on short sequences. Short subsequences (from 50 to 400 amino acids) were randomly sampled from the test dataset, and classification performance was compared among the nine methods listed in Table 2. Figure 6 compares % identifications by DA and other methods. With test sequences with 50 or 75 amino acids in length, we could still obtain higher than 70 % of positive identification by DA methods. In particular, nonparametric DA methods (KNN and nonparametric LDA) showed true positive rates higher than 80 % if sequences were longer than 75 amino acids. On the contrary, Pfam, for example, could identify fewer than 50 % of GPCR sequences when their length was 50 amino acids. Both PROSITE/pattern and PRINTS showed the lowest performance especially when sequences are short (300 amino acids or shorter). These observations were again consistent with the short patterns or fingerprints these classification methods use.

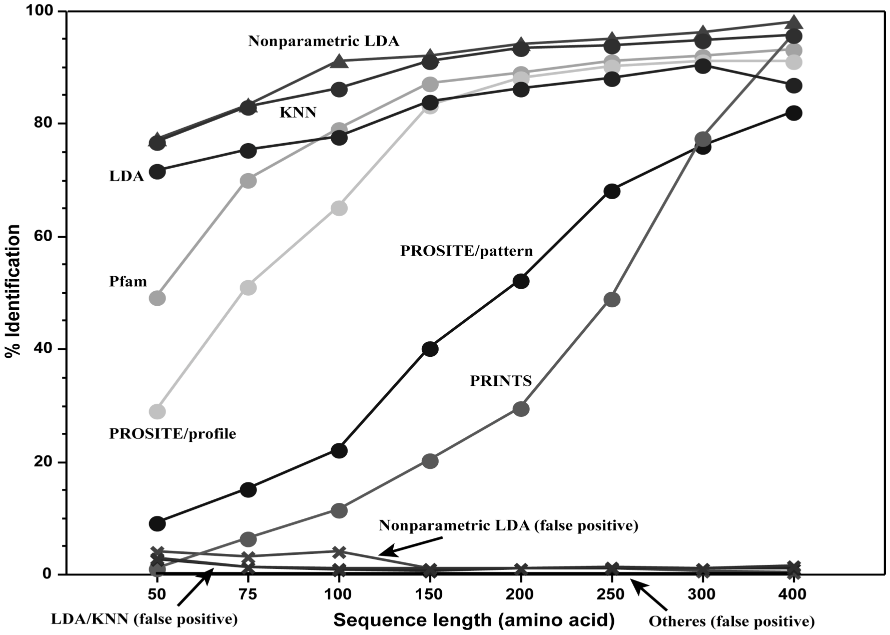


Figure 6. Performance comparison among classification methods against short protein sequences.

3. DEVELOPMENT OF A HIERARCHICAL CLASSIFICATION ALGORITHM

Our study showed that each protein classification method has different strength and weakness. All DA methods we examined (parametric or nonparametric) showed better performance even when protein sequences are extremely short. On the other hand, in general, DA methods had much higher false positive rates. We should note that we simply relied on SWISS-PROT annotations to identify GPCRs in this study. Therefore, it is possible that there are some miss-identifications. On the other hand, currently used methods (represented by PROSTIE, Pfam, and PRINTS in this study) have very low false positive rates. However, they perform best when properties of query and training datasets are consistent, as shown in their weak performance against partial sequences. Therefore, these methods are not likely to identify new protein sequences if they are not closely related to any existing protein family or any existing protein family member. These methods also rely on the quality of multiple alignments among training sequences.

The ideal protein classification method should have reasonably low false positive rates but needs to be sufficiently flexible, so new types of proteins can be still identified or classified. In order to realize such optimal means of protein classification, we are currently developing an integrated algorithm that compliments weakness of various methods by combining DA and other methods systematically and effectively.

The flowchart shown in the next page shows a simple example of such hierarchical algorithm to identify potentially new GPCR sequences. In this hierarchical algorithm, more "non-specific" DA methods are first used to identify any possible candidate for GPCR proteins and discriminating them from "least likely to be GPCR" proteins. In the next step, other more stringent methods (*e.g.*, PROSITE, Pfam, and PRINTS) can be used to filter out "more likely to be GPCR" data. The remaining dataset could contain both actual "false positives" and also some new members of the GPCR family. Another level of DA or other clustering methods could be used to select such possible candidates and to perform more detailed classification.

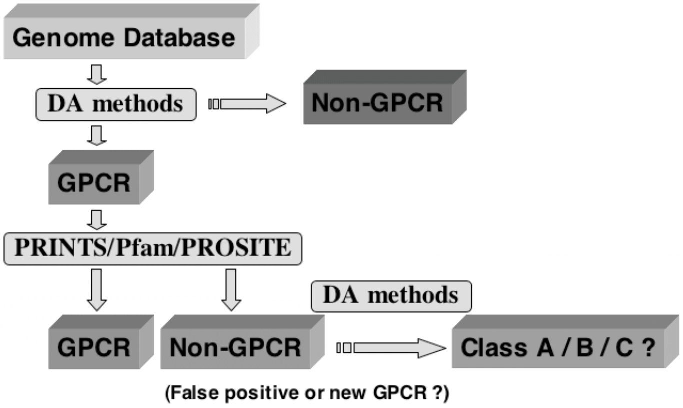


Figure 7. A hierarchical protein classification algorithm.

Further examinations of various DA and other multivariate methods are required to identify which "non-specific" methods should be incorporated in this algorithm. It is also possible to create another hierarchical level among these "non-specific" methods to lower the false positive rates.

Our goal is to develop an integrated hierarchical algorithm that can take advantage from various protein classification methods. This algorithm can be applied for any kind of protein families, and incorporating DA and other flexible methods, we can apply this algorithm even for partial or short sequences as found in EST databases. It will be also useful to identify particular protein coding sequences from short fragments (*e.g.*, exons) from genomic data.

4. REFERENCES

Altschul, S. F., Gish, W., Miller, W., Myers, E. W., and Lipman, D. J., 1990, Basic local alignment search tool, *J Mol Biol.* **215**:403; http://www.ncbi.nlm.nih.gov/BLAST/.

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; http://www.ncbi.nlm.nih.gov/BLAST/.

Attwood, T. K., Blythe, M., Flower, D. R., Gaulton, A., Mabey, J. E., Maudling, N., McGregor, L., Mitchell, A., Moulton, G., Paine, K., and Scordis, P., 2002, PRINTS and PRINTS-S shed light on protein ancestry, *Nucleic Acids Res.* **30**:239;

http://www.bioinf.man.ac.uk/dbbrowser/PRINTS/.

Bateman, A., Birney, E., Cerruti, L., Durbin, R., Etwiller, L., Eddy, S. R., Griffiths-Jones, S., Howe, K. L., Marshall, M., and Sonnhammer, E. L. L., 2002, The Pfam protein families database, *Nucleic Acids Res.* **30**:276; http://pfam.wustl.edu/index.html.

Falquet, L., Pagni, M., Bucher, P., Hulo, N., Sigrist, C. J., Hofmann, K., and Bairoch, A., 2002, The PROSITE database, its status in 2002. *Nucleic Acids Res.* **30**:235; http://www.expasy.ch/prosite.

Kim, J., Moriyama, E. N., Warr, C. G., Clyne, P. J., and Carlson, J. R., 2000, Identification of novel multi-transmembrane proteins from genomic databases using quasi-periodic structural properties., *Bioinformatics.* **16**:767.

Letunic, I., Goodstadt, L., Dickens, N. J., Doerks, T., Schultz, J., Mott, R., Ciccarelli, F., Copley, R. R., Ponting, C. P., and Bork, P., 2002, Recent improvements to the SMART domain-based sequence annotation resource, *Nucleic Acids Res.* **30**:242; http://smart.embl-heidelberg.de/.

Engelman, D. M., Steitz, T. A., Goldman, A., 1986, Identifying nonpolar transbilayer helices in amino acid sequences of membrane proteins, *Annu. Rev. Biophys. Biophys. Chem.* **15**:321.

Venables, W. N., and Ripley, B. D., 2002, *Modern Applied Statistics with S.* Fourth Edition, Springer, New York.