EXCLI Journal 2010;9:133-140– ISSN 1611-2156
Received: July 14, 2010, accepted: September 23, 2010, published: October 04, 2010

Original article:

FAST PROTEIN CLASSIFICATION BY USING THE MOST SIGNIFICANT PAIRS

Essam Al-Daoud

Computer Science Department, Faculty of Science and Information Technology, Zarka Private University, Zarka, Jordan, Telephone: 962-796680005
e-mail: essamdz@zpu.edu.jo

ABSTRACT

This study introduces a new approach to speed up the protein classification process. The basic idea is rewriting the sequences of each family by using the most significant pairs, where the total number of the pairs that can be appeared in the protein sequences is 400 different pairs. The sequence length could be reduced to 0.86, 0.91 and 0.95 by using the most 100, 200 and 300 significant pairs, respectively. The average time reduction is 0.53 %, 0.33 % and 0.22 % for 100, 200, and 300 pairs, respectively. In the three cases the suggested procedure can be adopted to speed up the testing time. However to get identical classification rate to the previous profile HMM, 300 pairs at least must be used.

Keywords: Hidden Markov, multi alignment, G-Protein Coupled Receptor, significant pairs

INTRODUCTION

Proteins have many important structures and functions, e.g., enzymes, antibodies, hormones, transport, molecules, hair, skin, muscle, tendons, cartilage, claws, nails, horns, hooves and feathers are all made of proteins. They are built from 20 amino acids and have a common chemical core of two functional groups, a carboxylic acid and an amino group. The enormous diversity of protein is due to the many ways in which amino acids can be combined, e.g., 20^{500} sequences can be formed by using a chain consists of 500 amino acids. Protein sequence classification is an important problem in biological sciences for annotation new protein sequence, detecting close evolutionary relationships among sequences or discovering new drug (Sabouri et al., 2010). In this paper, we should distinguish between four terms: family, domain, repeat and motifs. The family is a collection of related proteins, but the domain is a structural unit which can be found in multiple protein contexts, while the repeat is a short unit which is unstable in isolation but forms a stable structure when multiple copies are present, and the motifs is short unit found outside globular domains. Two main problems that make protein classification a difficult task: the first problem is that the number of possible protein is extremely large; the second problem is that the stability of the protein is not fully understood. Moreover, exploring the proteins by using the laboratory techniques such as spectroscopy and far ultraviolet are time consuming and expensive. On the other hand, much progress is being made by using the computational and statistical aspects. Profile Hidden Markov models (Profile-HMMs) are a widely used probabilistic modeling method for protein families that provides a probabilistic measurement (score) of how well an unknown sequence fits to a family (Rahman et al., 2009). A good target to test the performance of a new classification

EXCLI Journal 2010;9:133-140– ISSN 1611-2156
Received: July 14, 2010, accepted: September 23, 2010, published: October 04, 2010

method is G-Protein Coupled Receptor (GPCR) for two reasons: first GPCRs are a large and diverse family of proteins with over 360 identified subfamilies. Second GPCRs are transmembrane proteins which initiate via G-proteins some of the important signaling pathways in a cell and are involved in various physiological processes. Thus, computational prediction and classification of GPCRs can supply significant information for the development of novel drugs in pharmaceutical industry.

Various approaches have been developed for solving the protein classification problem. Most of them are based on appropriately modeling protein families, either directly or indirectly. Direct modeling techniques use a set of sequences to build and train a model that characterizes the family of interest e.g., Profile Hidden Markov models. Indirect techniques use direct models as a preprocessing tool in order to extract useful sequence features. In this way, sequences of variable length are transformed into fixed-length input vectors that are subsequently used for training discriminative models, such as BLAST, Support Vector Machines, decision trees, and Naïve Bayes classifier, and neural networks (Blekas et al., 2005; Gao & Wang, 2006). Papasaikas et al. (2003) presented a new method based on a probabilistic approach that exploits highly discriminative profile Hidden Markov Models, excised from low entropy regions of multiple sequence alignments, to derive potent family signatures. A best-guess family membership is depicted, allowing GPCRs' classification at a family level, solely using primary structure information. Blekas et al. (2005) considered two alternative ways for identifying the motifs to be used for feature generation and provide a comparative evaluation of the two schemes. They also evaluate the impact of the incorporation of background features (2-grams) on the performance of the neural system. Moriyama and Kim (2005) introduced a set of new methods that can classify protein family sharing very weak similarity. They described an algorithm that combines strengths from various protein classification methods to obtain an optimum power for protein classifications. Ergüner et al. (2006) developed a novel method for obtaining class specific features, based on the existence of activating ligand specific patterns, and utilized for a majority voting classification. Exploiting the fact there is a non-promiscuous relationship between the specific binding of GPCRs into their ligands and their functional classification. Benkrid et al. (2008) described the acceleration of the Viterbi decoding process by means of parallelizing the algorithm and mapping it to a systolic array. The concurrency of the array's processing elements is realized by implementing the engine on off-the-self FPGA hardware. Mathkour et al. (2010) proposed an integrated approach for the prediction of tri-nucleotide base patterns in DNA strands leading to transcription of peptide regions in genomic sequences. The approach comprises of preprocessing of data, transcription engine and post processing of output. The task has been carried out using series of filters that purify the raw data and assign weights to bases for further feeding to central engine.

MATERIALS AND METHODS

Data collection and preprocessing

In order to test the suggested procedures, G-Protein Coupled Receptor (GPCR) classification task was selected. GPCR is one of the current focus areas of pharmaceutical research. In addition to the biological importance of their functional roles, their interaction with more than 50 % of prescription drugs have led GPCRs to be an excellent potential therapeutic target class for drug design. GPCR is organized into a hierarchy of classes, Level 1, Level 2 and types. GPCR Level 1 consists of six different families: Class A, Class B, Class C, Class D, Class E, and Class Z.

EXCLI Journal 2010;9:133-140– ISSN 1611-2156
Received: July 14, 2010, accepted: September 23, 2010, published: October 04, 2010

Table 1: Level 1 of GPCR families and their Pfam code

Class	Name	Code	Number	Average Length
Class A	7tm_1	PF00001	1160	216.2
Class B	7tm_2	PF00002	1317	219.1
Class C	7tm_3	PF00003	941	218.3
Class D	STE3	PF02076	239	233.8
Class E	Dicty_CAR	PF05462	93	208.3
Class Z	Bac_rhodopsin	PF01036	1000	145.8
Total			4750	203.32

Table 2: Random protein families with variant length

Family Name	Number	Average Length	Description
Glucan_synthase	319	600.80	1,3-beta-glucan synthase component
Herpes_MCP	116	1090.10	Herpes virus major capsid protein
HOOK	150	517.90	HOOK protein
IpgD	70	507.50	Enterobacterial virulence protein IpgD
Med5	69	706.20	Mediator complex subunit Med5
Med23	68	750.10	Mediator complex subunit 23
Nrap	198	642.00	Nrap protein
OPT	1097	533.30	OPT oligopeptide transporter protein
Penicil amidase	694	686.50	Penicillin amidase
Reovirus_L2	33	1142.60	Reovirus core-spike protein lambda-2
Total	2814	624.5	

Table 3: The most 15 significant pairs

Pair	LL	EL	DE	EV	LF	IG	VP	FD	IF	EP	FI	NL	PE	FE	KE
Freq.	15	13	12	10	10	9	9	8	8	7	7	7	7	6	6

Each family consists of many subfamilies, for example Class A consists of amine, peptide, hormone, and rhodopsin. Amine protein is classified into seven subfamilies: muscarinic, adrenoceptors, dopamine, histamine, serotonin, octopamine, and trace amine.

For our experimental study three real datasets were sampled. The first dataset is GPCR Level 1 families from www.gpcr.org/, which is a large collection of G-protein coupled receptors families, summarized in Table 1. The second dataset is a random family from http://pfam.sanger.ac.uk/, summarized in Table 2. The third dataset consists of 4000 random protein sequences collected randomly from www.ncbi.nlm.nih.gov.

Profile HMM

A general Markov model is stochastic process in which S_i depends on the past (past time, past locations or past states) random variables $\{S_j\}$ where $j=i-1, i-2,...,1$:

$$Pr(S_n \backslash S_{n-1}, S_{n-2}, \ldots, S_1)$$

If n is large, then we have to collect a big number of the observation. Therefore, we have to approximate the general Markov model by using the first-order Markov model such as:

$$Pr(S_n \backslash S_{n-1}, S_{n-2}, \ldots, S_1) \approx Pr(S_n \backslash S_{n-1})$$

And the probability of a certain sequence $\{S_1, S_2, \ldots, S_n\}$ is

EXCLI Journal 2010;9:133-140– ISSN 1611-2156
Received: July 14, 2010, accepted: September 23, 2010, published: October 04, 2010

$$Pr\{S_1,S_2,...,S_n\} = \prod_{i=1}^{n} Pr(S_1|S_{i-1})$$

Unfortunately, the Markov model is not good enough to express the stochastic features of bioinformatic problems; the reason is that the relative frequencies of the state transitions in the data are used as state transition probabilities, which are different from real sequences. Hidden Markov Model (HMM) is more popular in modeling bioinformatic problems. HMM is represented by a set of parameters $\Theta = \{A, B, \pi\}$:

S: The set of (hidden) states $\{S_1, S_2, ...,S_n\}$.
O: The set of observations $\{O_1, O_2, ...,O_m\}$.
π (i): The initial state probability of the state
S_i
A: The probability transition matrix a_{ij}, where a_{ij}=Pr(state S_i at time t+l| state S_i at time t)
B: The probability output matrix b_{jk}, where b_{jk}=Pr(producing the observation O_k at time t| in state S_j at time t).

Three main issues are associated with HMM:

1- **Learning Problem:** Given some training observation sequences $O=\{O_1, O_2, ...,O_m\}$, and a HMM structure, determine HMM parameters $\Theta = \{A, B, \pi\}$ that best fit training data, and maximize $Pr(O|\Theta)$. Unfortunately, there is no feasible direct (optimal) solution, **Baum-Welch** algorithm is a good approximation solution for this problem.

2- **Evaluation Problem:** Given HMM $\Theta=\{A, B, \pi\}$ and the observation sequences $O=\{O_1, O_2, ..., O_m\}$, calculate the probability that model Θ has generated sequence O. The complexity of finding the all paths is $O(n^m)$ which is unfeasible. Therefore, an approximation solution can be calculated by using **Forward-Backward** algorithm.

3- **Decoding Problem:** Given the HMM $=\{A, B, \pi\}$ and the observation sequences $O=\{O_1, O_2, ...,O_m\}$, calculate the most likely sequence of hidden states S_i that produce this observation sequence. **Viterbi** algorithm is an efficient solution for this problem.

Profile HMM is a special HMM structure, a typical profile HMM architecture is shown in Figure 1. In addition to start and end state, there are three classes of states: the match states, the delete states and the insert states with $S=\{start, m_1, m_2,..., m_n, i_1, i_2,..., i_{n+1}, d_1, d_2,...d_n, end\}$, where n is the length of the model, typically equal to the average length of the sequences in the family. Once a profile HMM has been successfully trained on a family of sequences, it represents a model of the entire family and can be used for recognition, searching, multi-alignment, or classification tasks.

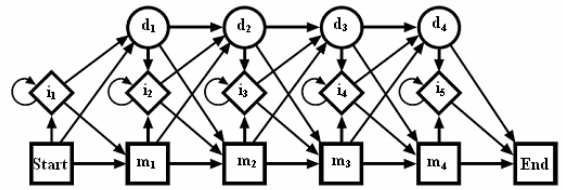

Figure 1: Profile HMM architecture

The proposed method

A protein sequence is made from various combinations of 20 amino acids, Let *Seq* be a protein sequence then $Seq \in \Sigma^n$, where Σ={A, C, D, E, F, G, H, I, K, L, M, N, P, Q, R, S, T, V, W, Y}. However, we can represent *Seq* as following:

$$Seq \in \begin{cases} (\Sigma\times\Sigma)^{n/2} & n \ is \ even \\ (\Sigma\times\Sigma)^{(n-1)/2}\times\Sigma & n \ is \ odd \end{cases}$$

To reduce the length of the protein sequences, we can rewrite the sequences of each family by using the most significant pairs. The total number of the pairs that can be appeared in the protein sequences is 400 different pairs i.e., $\Sigma x\Sigma$={AA, AC, AD, AE, AF, ..., CA, CC, CD, ..., YY} and $|\Sigma x\Sigma|$= 400. The significant pairs can be found by counting the frequencies of the pairs in the whole family. For example; consider the first 10 seed sequences of the family AgrD (PF05931):

EXCLI Journal 2010;9:133-140– ISSN 1611-2156
Received: July 14, 2010, accepted: September 23, 2010, published: October 04, 2010

Seq1:
MKLVNLLLSSTTSFLQMVGNRQKA
KTCTVLYDEPEVPKELTQELE
Seq2:
KKLLNKVIELLVDFFNSIGYRAAYIN
CDFLLDEAEVPKELTQLHE
Seq3:
NTLFNLFFDFITGILKNIGNIAAYSTC
DFIMDEVEVPKELTQLHE
Seq4:
TVLVDLIIKLFTFLLQSIGTIASFTPCT
TYFDEPEVPEELTNAK
Seq5:
MQIFDLLFKVISFIFEKIGFLAGYRTC
NTYFDEPEVPKELFETYQ
Seq6:
MQIINLLFKVITAVFEKIGFIAGYSTC
SYYFDEPEVPKELLEIYK
Seq7:
MRILEVLFNLITNLFQSIGTFARIPTS
TGFFDEPEIPAELLEEEK
Seq8:
MDILNGIFKFFAFIFEQIGNIAKYNPC
VGYFDEPEVPSELLDEQK
Seq9:
MELLNGIFKLFAFIFEKIGNLA-
KYYPCFGYFDESEVPQELLDEDK
Seq10:
MDLLNGIFKLFAFIFEKIGNLAKYNP
CLGFLDEPTVPKELLEEDK

The most 15 significant pairs are given in Table 3. If we re-encode the first 3 sequences by using the most 15 significant pairs then the sequences can be written as following:

Seq1: NLLLDEPEVPKELEL
Seq2: LLELLIGLLDEEVPKEL
Seq3: LFNLFFDFIIGFIDEVEVPKEL

While if we use the most 100 significant pairs then the sequences can be written as following:

Seq1:
KLVNLLLSTTSFLQVGNQKAKTCTV
LDEPEVPKELTQELE
Seq2:
KLLNKVIELLVDFFNSIGYRAAYINC
DFLLDEAEVPKELTQLHE
Seq3:
NTLFNLFFDFITGILNIGNIAAYSTCD
FIMDEVEVPKELTQLHE

And if we use the most 200 significant pairs then the sequences can be written as following:

Seq1:
MKLVNLLLSSTTSFLQMVGNRQKA
KTCTVLYDEPEVPKELTQELE
Seq2:
KKLLNKVIELLVDFFNSIGYRAAYIN
CDFLLDEAEVPKELTQLHE
Seq3:
NTLFNLFFDFITGILKNIGNIAAYSTC
DFIMDEVEVPKELTQLHE

The average of the reduction by using the most 15, 100 and 200 significant pairs are 0.39, 0.86 and 0.91, respectively.

Let $ArrFreq_i^{\alpha} = \theta(\lambda_i^{seed}, \alpha)$ be an array contains the most significant α pairs of the seed sequences of the protein family λ_i (seed sequences are a set of sequences that have been manually checked by experts such as Pfam seed). Algorithm 1 illustrates the main steps to train the HMM by using the proposed frequency array. Algorithm 2 can be used to determine whether a protein sequence belongs to the family λ_i or not.

Algorithm 1: Train Profile HMM
Input: $ArrFreq_i^{\alpha}$ and λ_i.
Output: Trained Profile HMM
1- β = Encode (λ_i, $ArrFreq_i^{\alpha}$). // Encodes the family λ_i by using the pairs in $ArrFreq_i^{\alpha}$.
2- δ =MultiAlign(β). // Multi Alignment the sequences in β.
3- ψ = hmmprofestimate (size (δ), δ). // Estimates the HMM parameters for the aligned sequences in δ.

Algorithm 2: Test a sequence
Input: An unknown sequence *x*, HMM parameters ψ generated from the family λ_i and $ArrFreq_i^{\alpha}$.

EXCLI Journal 2010;9:133-140– ISSN 1611-2156
Received: July 14, 2010, accepted: September 23, 2010, published: October 04, 2010

Output: if x belongs to λ_i then "yes" else "no".

1- $\mu = \text{Encode}\ (x, ArrFreq_i^{\alpha})$. // Encodes the sequence x by using the pairs in $ArrFreq_i^{\alpha}$.

2- $v = \text{hmmprofalign}(\psi, x)$. // Returns the score for the optimal alignment of the query sequence. Scores are computed using log-odd ratios for emission probabilities.

3- If $v > O$ return "yes" else return "no".

RESULTS AND DISCUSSION

Several experiments were conducted to evaluate the proposed method. The classification accuracy was measured by counting the true positives (TP), false positives (FP), true negatives (TN) and false negatives (FN). The accuracy rate is defined as the proportion of correct predictions and is given by:

Table 4: Testing time and recognition rate of the family 7tm_1

	No.	Avg. Len.	α =100		α =200		α =300		Previous	
			T	S	T	S	T	S	T	S
Positive Sequences	1000	216.8	63	0.19	97.2	.28	95.1	0.32	95.1	0.41
Negative Random Families	2814	624.5	65	0.31	97	0.42	96.9	0.49	96.9	0.61
Negative Random Sequences	4000	1123	62	0.51	96.8	0.74	98	0.86	98	1.15
Negative GPCRs	3590	199.1	63	0.14	97	0.21	98	0.27	98	0.35
Avg. & Rate	11404	629.7	63.1	0.31	97	0.45	97.5	0.53	97.5	0.70

Table 5: Testing time and recognition rate of the family 7tm_2

	No.	Avg. Len.	α =100		α =200		α =300		Previous	
			T	S	T	S	T	S	T	S
Positive Sequences	1317	219.1	59	0.15	89	0.17	92.4	.20	92.4	.24
Negative Random Families	2814	624.5	60	0.33	95.7	0.52	96.5	0.6	96.5	.7
Negative Random Sequences	4000	1123	60	0.51	95.2	0.70	97	0.69	97	1.0
Negative GPCRs	3273	196.6	61	0.16	95	0.20	97	0.24	97	0.32
Avg. & Rate	11404	629.7	60.2	0.32	94.7	0.43	96.5	0.48	96.5	0.64

Table 6: Testing time and recognition rate of the family 7tm_3

	No.	Avg. Len.	α =100		α =200		α =300		Previous	
			T	S	T	S	T	S	T	S
Positive Sequences	941	218.3	60	0.16	92.1	0.24	93.9	0.3	93.9	0.40
Negative Random Families	2814	624.5	64	0.3	97	0.41	98.2	0.5	98.2	0.65
Negative Random Sequences	4000	1123	63	0.49	97	0.76	97.6	0.9	97.6	1.0
Negative GPCRs	3649	199.0	64	0.14	97	0.17	98.3	0.22	98.3	0.35
Avg. & Rate	11404	629.7	63.3	0.30	96.6	0.44	97.6	0.53	97.6	0.65

EXCLI Journal 2010;9:133-140– ISSN 1611-2156
Received: July 14, 2010, accepted: September 23, 2010, published: October 04, 2010

$$\text{Accuracy Rate} = \frac{TP + TN}{TP + TN + FP + FN}$$

TP is the number of actual GPCRs that are predicted as GPCRs, *FP* is the number of actual non-GPCRs that are predicted as GPCRs, *TN* is the number of actual non-GPCRs that are predicted as non-GPCRs, and *FN* is the number of actual GPCRs that are predicted as non-GPCRs. Profile HMM in Algorithm 1 was trained by using three different GPCR families i.e. 7tm_1, 7tm_2 and 7tm_3, where the seed size was 50 and three α values were used, i.e. α=100, α=200, α=300. The results were compared with the previous profile HMM approach. In tables 3-5 two main sets were used, the first were positive sequences which belong to the trained family and the second were negative sequences which are divided into three subsets, i.e. negative random families from Table 2, negative random sequences, and negative GPCRs which belong to the rest of GPCR. T denotes to true positive or true negative, S denotes to the testing time in seconds per sequence.

In Table 4, the seed sequences of the family 7tm_1 were used to train Algorithm 1 (50 sequences). 1000 positive sequences and 10404 negative sequences were used for testing. The positive sequences belong to the family 7tm_1. It can be noticed that the average time reduction are 0.55 %, 0.35 % and 0.24 % for α=100, α=200, α=300, respectively. In Table 5 the seed sequences from family 7tm_2 were used to train Algorithm 1 (50 sequences). 1317 positive sequences and 10087 negative sequences were used for testing. The average testing time reduction are 0.50 %, 0.32 % and 0.25 % for α=100, α=200, α=300, respectively. In Table 6 the seed sequences from family 7tm_3 were used to train algorithm 1 (50 sequences). 941 positive sequences and 10463 negative sequences were used for testing The average time reduction are 0.53 %, 0.32 % and 0.18 % for α=100, α=200, α=300, respectively. We can conclude that, in the three cases the suggested procedure can be adopted to speed up the testing time. However to get identical classification rate to the previous profile HMM, α=300 must be used.

CONCLUSION

A large amount of new protein sequences are being accumulated in various databases. An important task for researchers in bioinformatics is to classify these proteins in families based on their structural and functional properties. Although laboratory experiments are the most reliable, they are not cost and time effective. To automate the process, computation methods have been extensively used. In this study a new approach to speed up the protein classification process was introduced. The first step of the suggested procedure is counting the frequency of the amino acid pairs of a protein family, and then rewriting the seed sequences by using the most significant pairs, multi-alignment and estimating the profile-HMM parameters of the modified sequences. In order to classify a protein sequence, we must rewrite the test sequence by using the same pairs. The test sequence is classified according to the score of the profile-HMM. The suggested procedure can be adopted to speed up the testing time from 0.22 % - 0.53 % by using the most 100-300 significant pairs.

REFERENCES

Benkrid K, Velentzas P, Kasap S. A high performance reconfigurable core for motif searching using profile HMM. NASA/ESA Conference on Adaptive Hardware and Systems, AHS' 2008, pp 285-92. Noordwijk, 2008.

Blekas K, Fotiadis DI, Likas A. Motif-based protein sequence classification using neural networks. J Comput Biol 2005;12: 64-82.

Ergüner B, Erdoğan Ö, Sezerman U. Prediction and classification for GPCR sequences based on ligand specific features. In: Levi A et al. (Eds.): 21th International Symposium, ISCIS, Istanbul, Turkey, 2006 (pp 174-81). Berlin: Springer-Verlag (Lecture notes in computer science, Vol. 4263).

Moriyama EN, Kim J. Protein family classification with discriminant function analysis. In: Gustafson JP, Shoemaker R, Snape JW (Eds.): Genome exploitation. Data mining the genome (pp 121-32). New York: Springer Science + Business Media, 2005.

Gao Q-B, Wang Z-Z. Classification of G-protein coupled receptors at four levels. peds Protein Engin Design Select 2006;19:511–6.

Mathkour H, Ahmad M. An integrated approach for protein structure prediction using artificial neural network. In: 2nd International Conference on Computer Engineering and Applications (ICCEA), 2010, Bali Island, pp 484-8.

Papasaikas PK, Bagos PG, Litou ZI, Hamodrakas SJ. A novel method for GPCR recognition and family classification from sequence alone using signatures derived from profile hidden Markov models. SAR QSAR Environ Res 2003;14:413-20.

Rahman SA, Hussein ZAM, Bakar AA. Experimental study of different FSAs in classifying protein function. In: International Conference of Soft Computing and Pattern Recognition (SOCPAR), 2009 (pp 516-21). Malacca, 2009.

Sabouri A, Ardalan A, Shahidi-Nejad R. Prediction of protein secondary structure based on NMR chemical shift data using support vector machines. In: 12th International Conference on Computer Modelling and Simulation (UKSim), 2010, pp 201-5. Cambridge, 2010.