CABIOS

Vol. 12 no. 2 1996
Pages 109–118

Motif identification neural design for rapid and sensitive protein family search

Cathy H.Wu[1], Sheng Zhao, Hsi-Lien Chen, Chin-Ju Lo and Jerry McLarty

Abstract

A new method, the motif identification neural design (MOTIFIND), has been developed for rapid and sensitive protein family identification. The method is an extension of our previous gene classification artificial neural system and employs new designs to enhance the detection of distant relationships. The new designs include an n-gram term weighting algorithm for extracting local motif patterns, an enhanced n-gram method for extracting residues of long-range correlation, and integrated neural networks for combining global and motif sequence information. The system has been tested and compared with several existing methods using three protein families, the cytochrome c, cytochrome b and flavodoxin. Overall it achieves 100% sensitivity and > 99.6% specificity, an accuracy comparable to BLAST, but at a speed of ~20 times faster. The system is much more robust than the PROSITE search which is based on simple signature patterns. MOTIFIND also compares favorably with BLIMPS, the Hidden Markov Model and PROFILESEARCH in detecting fragmentary sequences lacking complete motif regions and in detecting distant relationships, especially for members of under-represented subgroups within a family. MOTIFIND may be generally applicable to other proteins and has the potential to become a full-scale database search and sequence analysis tool.

Introduction

As technology improves and molecular sequencing data accumulate nearly exponentially, progress in the Human Genome Project will depend increasingly on the development of advanced computational tools for rapid and accurate annotation of genomic sequences. Currently, a database search for sequence similarities is the most direct computational means of deciphering codes that connect molecular sequences with protein structure and function (Doolittle, 1990). There are good algorithms and mature software for database search and sequence analysis (Gribskov and Devereux, 1991), which may be based on pair-wise comparisons between the query sequence and sequences in the molecular database. Alternatively, a database search may be based on information derived from a family of related proteins. This includes methods that screen for motif patterns such as those cataloged in the PROSITE database (Bairoch and Bucher, 1994), the Profile method (Gribskov *et al.*, 1987), the hidden Markov model (HMM) (Krogh *et al.*, 1994; Eddy *et al.*, 1995), and the neural network classification method (Wu *et al.*, 1992; Wu, 1995).

With the accelerating growth of the molecular sequence databases, it is widely recognized that database searching against gene/protein families or motifs is an important strategy for efficient similarity searching (Altschul *et al.*, 1994). This is evidenced by the growing efforts in recent years for building second generation (or secondary value-added) databases that contain domains, motifs or patterns (Henikoff and Henikoff, 1991; Attwood *et al.*, 1994; Sonnhammer and Kahn, 1994). While several domain/motif databases are being compiled, it is important to develop database search methods that fully utilize the conserved structural and functional information embedded in those databases to enhance search sensitivity. In this paper we report a new motif identification neural design for rapid and sensitive protein family identification, and compare it to the current state of the art methods, including the BLAST database search (Altschul *et al.*, 1990), the PROSITE pattern search, the BLIMPS search of BLOCKS (Wallace and Henikoff, 1992), the HMM method, and the PROFILESEARCH (Gribskov *et al.*, 1989).

System and methods

Data sets

Three protein families of electron transferases, cytochrome c, cytochrome b and flavodoxin, were used to test the MOTIFIND system (Table I). The three families represent three different case studies. The cytochrome c family, grouped according to PROSITE pattern PS00190 and whose members share the same C-x-x-C-H heme-binding site signature, actually is a widely diversified family that consists of cytochrome c, c′, c1 to c6, c550 to c556, cytochrome f and reaction center cytochrome c. The family members vary greatly in their lengths, ranging from ~85 aa (amino acids) to > 300 aa. The size of subgroups

Department of Epidemiology/Biomathematics, The University of Texas Health Center at Tyler, Tyler, TX 75710, USA

[1]To whom correspondence and reprint requests should be addressed. Email: wu@jason.uthct.edu

within the family also varies, from the most abundant cytochrome c with 85 entries to small subgroups with < 5 entries. The cytochrome b family has two subgroups, the cytochrome b of around 380 aa in length and b6 of ~220 aa. The flavodoxin is a small, homogeneous family containing only 23 members.

The positive set consisted of all sequences of the protein family studied (Table I, Column 6). These included the sequences cataloged in the PROSITE database (Release 12.2, February 1995, compiled based on SWISSPROT database Release 29.0), as well as new sequences selected directly from the SWISSPROT database (Release 31.0, February 1995) (Bairoch and Boeckmann, 1994) by combinations of database sequence search/alignment, signature pattern search and manual examination of sequence annotation. The complete negative set contained all sequences in the SWISSPROT database that were non-members of the protein family studied (Table I, Column 7).

The training set for the neural network consisted of both positive (members of the protein family) and negative (non-members) patterns at a ratio of 1:2. Approximately two-thirds of the 'T' sequences (true positives) cataloged in PROSITE were chosen randomly as the positive training set. The negative training set were selected randomly from all non-members. Two prediction sets were used: the full-scale prediction set was the entire SWISSPROT database (Release 31.0), containing 43 470 sequences; and the small prediction set consisted of all positive patterns and randomly selected negative patterns. For cytochrome b and flavodoxin, all 'F' sequences (false positives) in the PROSITE database were also included as negative patterns. In MOTIFIND, the neural network training uses both full-length and motif sequences to obtain global and local information. The full-length sequences were directly taken from the SWISSPROT database. The motif sequences used to compute the n-gram weight factors were compiled by using our own string pattern-matching program to search for PROSITE signatures (Table I) and retrieve substrings in the BLOCKS format (Henikoff and Henikoff, 1991).

To ensure the validity of the MOTIFIND results, a three-fold cross-validation method was used. In this method, the positive training set for each family was randomly divided into three approximately equal sized sets. In each trial, one set was used for prediction and the remaining two sets for training. Since similar results were obtained using any of the three data sets for each protein family (results not shown), only the first dataset was used for all comparative studies described below.

Evaluation mechanism

The system performance was evaluated based on speed (CPU time) and predictive accuracy. Accuracy was measured in terms of both sensitivity (ability to detect true positives) and specificity (ability to avoid false positives) at different threshold values. Two types of scores were given to each query sequence after network prediction, the neural network scores and the probability (P) score. There were four neural network scores for each sequence, corresponding to the values of the four output units in the MOTIFIND network and these represent the full-length and motif neural network outputs for positive and negative classes, respectively (i.e. +F, +M, −F and −M scores, Figure 1). The P score was computed from the four neural network scores using a logistic regression function,

$$\log(P_{hit}/(1 - P_{hit})) = \alpha + \beta_1 O_1 + \beta_2 O_2 + \beta_3 O_3 + \beta_4 O_4 \quad (1)$$

where P_{hit} is the probability of hit, α, β_1 to β_4 are the regression parameters, and O_1, O_2, O_3 and O_4 are the MOTIFIND network scores. The logistic regression model is equivalent to a two-layered neural network (i.e. perceptron) with a logistic activation function (Sarle, 1994). We implemented the two-layer perceptron by adopting the same feed-forward and back-propagation functions (Wu *et al.*, 1992). The perceptron had four input units (which used O_1 to O_4 as inputs) and one output unit (whose score was P_{hit}), as well as a bias term (for computing the constant α).

A positive sequence is considered to be accurately predicted (i.e. true positive) if both the P score and the average (AVG) neural network score (i.e. the average of the +F and +M scores) are higher than certain threshold values (cut-off scores). Conversely, a negative (non-member) sequence is accurately predicted (i.e. true negative) if either score is lower than the threshold. Both neural network score and P score range between 0.0 (no match) and 1.0 (perfect match). Note that the P and AVG scores are related but different measures, because the former considers degrees of similarity to negative random sequences through the use of −F and −M scores. The SSEARCH program (version 1.7A, July 1994) (Smith and Waterman, 1981; Pearson, 1991) was used to determine the overall sequence similarity of a query sequence to the neural network training sequences.

Comparative studies

The MOTIFIND results were compared with several other existing methods. The BLAST search was performed using the improved version (version 1.4, October 1994) that adopted Sum statistics (Karlin and Altschul, 1993). The program was obtained from the NCBI FTP server (ncbi.nlm.nih.gov) and implemented on our DEC alpha

Table I. Data sets for neural network training and prediction

Protein Family	PROSITE number	Motif length[a]	Training set		Prediction Set		
			# Pos[b]	# Neg[c]	# Pos	# Neg F[c]	# Neg S[c]
Cytochrome C	PS00190	15	149	298	238	43 232	162
Cytochrome B	PS00192	41	86	172	151	43 319	101
Flavodoxin	PS00201	19	14	28	23	43 447	21

[a]The motif patterns, adopted from PROSITE signatures, are: x(8)-C-{CPWHF}-{CPWR}-C-H-{CFYW}-x (Cytochrome C); x(9)-[DENQ]-x(3)-G-[FYWM]-x-[LIVMF]-R-x(2)-H-x(13)-H-x(6) (Cytochrome B); and x(2)-[LIV]-[LIVFY]-[FY]-x-[ST]-x(2)-[AG]-x-T-x(3)-A-x(2)-[LIV] (Flavodoxin).
[b]# Pos, number of positive patterns. The 238 cytochrome c members were 230 (i.e., 225 'T' + 4'N' + 1 'P') sequences cataloged in PROSITE (release 12.2) and eight newly-identified members, including six 'T' sequences (C551_ECTHL, CY32_DESDN, CYC6_MONBR, CYC_EMENI, CYSD_CHRVI, YHJA_ECOLI) and two 'N' patterns (C553_PARDE, CY1_EUGGR). The cytochrome b family included 136 (131 'T' + 2 'N' + 2 'P' + 1 '?') PROSITE entries, and 14 new 'T' sequences (CYB_BALAC, CYB_BALBN, CYB_BALBO, CYB_BALED, CYB_BALGL, CYB_BALMU, CYB_BALMY, CYB_CAPMR, CYB_DIDMA, CYB_ESCGI, CYB_HALGR, CYB_LEPWE, CYB_MEGNO, CYB_PHYCA), and one new 'N' sequence (CYB_SULAC). The flavodoxin members were 22 (21 'T' + 1 'P') PROSITE entries and one new 'N' sequence (FLAW_ECOLI). The PROSITE codes are: 'T' (true positive), 'N' (false negative containing degenerate signature), 'P' (false negative lacking complete signature, usually fragmentary), '?' (uncertain), and 'F' (false positive).
[c]# Neg, Number of negative patterns; Neg F, Negative patterns in the full-scale prediction set, i.e. all negative sequences in the entire SWISSPROT database; Neg S = Negative patterns in the small-scale prediction set.

workstation running on OSF/1 operating system. The same training set (containing both positive and negative sequences) and prediction set used in MOTIFIND (Table I) were used as BLAST database and query sequences, respectively. The negative sequences were included as database entries because they provided much better class separations for BLAST (results not shown). The result reported was based on the probability score of the first-hit.

The PROSITE search was performed by using our own string pattern-matching program to search for PROSITE signatures. The results obtained with our pattern-matching program using the SWISSPROT database Release 29.0 were identical to those cataloged in the PROSITE database (Release 12.2).

The BLIMPS search involved BLOCKS building and search. To obtain the BLOCKS, the training sets (containing only positive sequences) were sent directly to the BLOCKMAKER (version 1.11, June 1994) Email server (blockmaker@howards.fhcrc.org) (Henikoff *et al.*, 1995). The individual BLOCKS were then used to search the prediction set with BLIMPS (version 2.2 A, May 1994) obtained from the NCBI server, using default amino acid frequency. The results presented were obtained by using the Gibb BLOCKS of 10 aa, 53 aa and 20 aa, respectively, for the cytochrome c, cytochrome b and flavodoxin families.

The HMMER program (ftp://genome.wustl.edu/pub/eddy/) (version 1.8, April 1995) was used to run the HMM method. It involved model building from aligned training sequences (positive sequences in the neural network training sets) using the HMMB program, and Smith-Waterman database search against the model using the HMMSW program. To improve the detection of under-represented family members, the maximum discrimination HMM (Eddy *et al.*, 1995) was studied (i.e. the HMMB program with '-d' option for maximum discrimination instead of maximum likelihood), using training sequences pre-aligned by the CLUSTALW program (ftp.bio.indiana.edu) (version 1.4, September 1994) (Thompson *et al.*, 1994). Both global and motif models were built for cytochrome c, and only global models were used for the more homogeneous families of cytochrome b and flavodoxin. The global and motif models were built with the same full-length and motif training sequences used in MOTIFIND. The lengths of the resultant models were 129 aa, 15 aa, 381 aa and 172 aa, respectively, for the cytochrome c global and motif models, cytochrome b global model, and flavodoxin global model.

The profile method was conducted using the GCG Wisconsin Sequence Analysis Package (version 8.0, September 1994), running on a VAX mainframe. There were three procedures: multiple sequence alignment of training sequences using PILEUP program, profile generation from the resulting alignment using PROFILEMAKE program, and database search against each profile using PROFILESEARCH program. Default parameters were used in all the programs. Profiles used in the comparative studies and their lengths were: motif profile for cytochrome c (15 aa), global and motif profiles for cytochrome b (505 and 41 aa) and global profile for flavodoxin (186 aa). All profiles were made from the same full-length and motif sequences used in MOTIFIND, except that only 100 sequences were selected for cytochrome c because of the size limitation of the PROFILEMAKE program.

Algorithm

There are two basic design concepts underlying the new

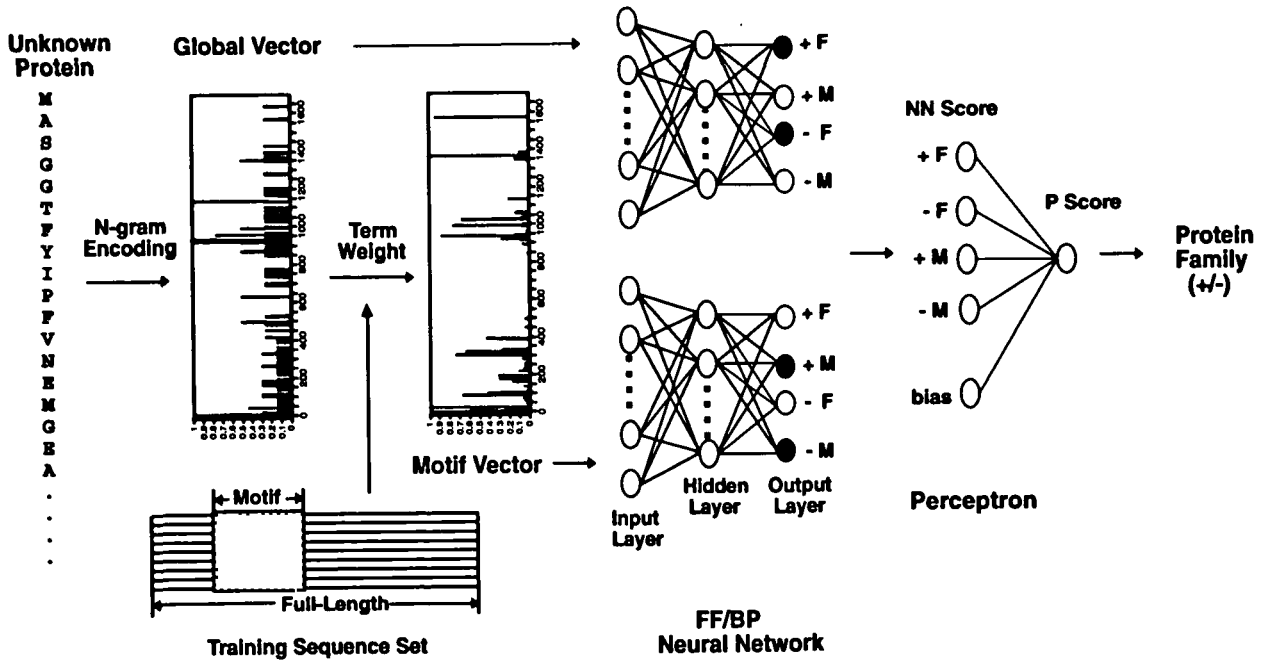

Fig. 1. MOTIFIND for rapid and sensitive protein family identification. Each sequence string is converted into two separate input vectors of real numbers (i.e. global and motif vectors) using an n-gram method to encode global sequence similarity and a term weighting method to extract motif information. Each of the two vectors is then mapped to the appropriate output units of a single three-layered, feed-forward, back-propagation (FF/BP) neural network, according to information embedded in the neural interconnections after network training. The four neural network scores corresponding to the full-length (F) and motif (M) output units for positive (+) and negative (−) classes are used to compute the probability (P) score using a two-layered neural network (Perceptron).

search method: (i) a fast one-step family identification that replaces pair-wise sequence comparisons of high computational cost; and (ii) the combination of global sequence similarity with conserved family information embedded in motif patterns to improve search accuracy. While we used the first design concept in our previous gene classification artificial neural system (GenCANS) (reviewed in Wu, 1995), we introduce three new designs to implement the second concept, an n-gram term weighting algorithm for extracting local motif patterns, an enhanced n-gram method for extracting residues of long-range correlation, and integrated neural networks for combining global and motif sequence information.

As depicted in Figure 1, the MOTIFIND search involves two steps, a sequence encoding step to convert protein sequences into neural network input vectors, and a neural network classification step to map input vectors to appropriate protein families. The sequence encoding schema involves an n-gram hashing function that extracts and counts the occurrences of patterns (terms) of n consecutive residues (i.e. a sliding window of size n) from a sequence string (Wu *et al.*, 1992). Unlike the FASTA method (Pearson and Lipman, 1988), which also uses n-grams (k-tuples), our search method uses the counts, not the positions, of the n-gram terms along the sequence. Therefore, our method is length-invariant, provides certain insertion/deletion invariance, and does not require the laborious sequence alignments of many other database search methods. In the encoding, each unit of the neural input vector represents an n-gram term, thus, the size of the input vector is m^n, where m is the size of the alphabet and n is the length of the n-gram. The original sequence string can be represented by different alphabet sets in the encoding, including the 20-letter amino acids and the six-letter exchange groups derived from the PAM (accepted point mutation) matrix. Different exchange groups can also be defined for different protein families to emphasize the conservative replacement unique for the family.

N-gram term weighting

A new n-gram term weighting method is used to extract conserved family information from motif sequences by multiplying each n-gram term with its weight factor. The weight factor is calculated by dividing the total n-gram counts in all motif sequences (term frequency) with total n-gram counts in all full-length sequences of the training set (inverse set frequency), as in:

$$W_k = \Sigma_i M_{ik} / \Sigma_i F_{ik} \quad (2)$$

where W_k is the weight factor for the k-th n-gram term in the input vector, and F_{ik} and M_{ik} are total counts of the k-th n-gram term in the i-th sequence of the full-length sequence set and motif set, respectively. The equation reflects that the n-gram terms of high weights are both conserved (present in all training sequences) with high term frequency, and unique (present in motif regions only) with high inverse set frequency. Also note that motif terms used by members of over-represented subgroups will not receive higher weights because the set frequency (term usage in full-length sequence) is considered.

Enhanced n-gram method

In addition to the global and motif term vectors described above, other specific n-gram terms that are highly conserved can also be included as additional input units to the neural networks. These highly conserved terms have strong signals and serve as effective indexes for family identification. The terms may involve neighboring residues, or may involve residues that are far apart in the linear sequence, as frequently found in structural motifs. In the original n-gram method, neighboring letters (with a distance of zero) are extracted. The enhanced method, on the other hand, uses terms of various distances to extract long-range correlation of amino acid residues, thereby, encode some of the most important positional information that are absent in the original n-gram encoding.

Integrated neural networks

The neural network classification employs three-layered, feed-forward, back-propagation networks (Wu *et al.*, 1992). As a technique for computational analysis, neural network technology has been applied to many studies involving sequence data analysis (Hirst and Sternberg, 1992), such as protein structure prediction, identification of protein-coding sequences, and prediction of promoter sequences. In this study, we use an integrated neural network design in which each protein family is represented by an individual neural network with multiple output units, one for each classification parameter.

Implementation

The system has been coded with C programs and implemented on the Cray supercomputer of the University of Texas System and a DEC alpha workstation, using a program structure similar to GenCANS (Wu, 1995). The system software has three components: a preprocessor to create the training and prediction patterns from input sequence files, a neural network program to classify input patterns, and a postprocessor to perform statistical analysis and summarize classification results.

The particular n-gram method used concatenated bi-grams of amino acids and tetra-grams of exchange groups, and resulted to a vector size of 1696 (i.e. $20^2 + 6^4$). The six-letter exchange groups were: {MILV}, {FYW}, {STPAG}, {QDENRK}, {H}, {C} for cytochrome c; {MILV}, (FYW}, {STPAC}, {QDENK}, {HR}, {G} for cytochrome b; and {MILV}, {FYW}, {STPAG}, {QDEN}, {HRK}, {C} for flavodoxin.

Two separate input vectors were generated from each sequence, a global vector and a motif vector (Figure 1). The global vector contained counts of the n-gram terms from the full-length sequence, scaled between 0 and 1; whereas the motif vector had counts multiplied with weight factors before scaling. The final input vectors used for the network training and prediction were generated by concatenating the specific terms (with k units, k may be varied for different families) to both global and motif term vectors. Thus, each of the two input vectors (global and motif) had a size of 1696 + k. The neural networks used to represent the cytochrome c, cytochrome b and flavodoxin families had 1696, 1698, and 1696 input units, respectively. The two additional input units in cytochrome b network were used to represent the two overlapping histidine heme ligands in the cytochrome b family, R–x(2)–H, H-x(13)–H, R–x(4)–H and H–x(14)–H, by using terms RH2 (i.e. RH n-gram with distance of 2) and RH4, and terms HH13 and HH14. The output layer had four units, representing two parameters (global and motif) for two classes (positive and negative sets). The hidden size was determined heuristically to be 20. Therefore, the final network architecture was (1696 + k) × 20 × 4. Other network parameters included: random initial weights of −0.3 to 0.3, a back-propagation learning factor of 0.3, a momentum term of 0.2, a constant bias term of −1.0, and an error threshold of 0.01.

Results

MOTIFIND performance

Table II shows that MOTIFIND achieved 100% sensitivity and >99.6% specificity in a full-scale SWISSPROT database search for all three protein families studied. There are several factors that may affect the predictive accuracy of a given sequence: (i) the degree of overall sequence similarity, (ii) the sequence length, (iii) the prevalence of the sequence in the family, and (iv) the existence of motif regions. MOTIFIND is capable of identifying not only full-length, closely related sequences, but also distantly related sequences, fragmentary sequences, and sequences of under-represented groups within the family. Close inspection of sequence patterns reveals that MOTIFIND can detect with high scores the

Table II. Comparisons of MOTIFIND with other database search and family identification methods in terms of speed and accuracy

Search method	CPU time[a]	Threshold[b]	Sensitivity[c]		Specificity F[c]		Specificity S[c]	
			(%)	true+	(%)	false+	(%)	false+
A. Cytochrome C								
MOTIFIND	984	0.43 / 0.30	100.00	238	99.62	166	100.00	0
		0.91 / 0.92	96.22	229	99.99	5[d]	100.00	0
BLAST	35116	6.7 E-03	100.00	238	99.08	396	99.38	1
		4.8 E-26	91.60	218	99.98	9	100.00	0
PROSITE	27	n/a	97.06	231	99.45	237	97.53	4
BLIMPS	172	287	99.58	237	98.49	653	97.53	4
		370	78.15	186	99.94	24	100.00	0
HMM Motif	425	−1.39	99.58	237	98.42	685	96.91	5
		4.90	87.39	208	99.98	9	100.00	0
HMM Global	4036	−5.56	97.06	231	93.29	2901	93.83	10
		4.14	86.97	207	99.98	9	100.00	0
PROFILE Motif	–[e]	4.02	99.16	236	–	–	96.30	6
B. Cytochrome B								
MOTIFIND	1452	0.43 / 0.46	100.00	151	99.95	23	100.00	0
BLAST	24597	1.4 E-22	100.00	151	99.99	3[f]	100.00	0
PROSITE	33	n/a	96.69	146	100.00	1	99.01	1[g]
BLIMPS	756	948	98.68	149	99.86	60	100.00	0
HMM Global	12183	34.64	99.34	150	99.97	13[f]	100.00	0
PROFILE Motif	–	11.83	98.01	148	–	–	100.00	0
PROFILE Global	–	43.70	98.68	149	–	–	88.12	12
C. Flavodoxin								
MOTIFIND	1019	0.48 / 0.89	100.00	23	99.99	5	100.00	0
BLAST	21411	9.0 E-05	100.00	23	99.95	23[h]	100.00	0
PROSITE	34	n/a	91.30	21	99.99	5	76.19	5[g]
BLIMPS	265	676	100.00	23	99.99	6[h]	100.00	0
HMM Global	5630	2.65	100.00	23	99.96	18[h]	95.24	1
PROFILE Global	–	31.31	91.30	21	–	–	95.24	1

[a]The time shown is the total CPU seconds required on a DEC alpha workstation to process the entire prediction set of 43 470 sequences.
[b]The results are shown at different cut-off scores in order to maximize the sensitivity or specificity of each given method. The threshold values shown for MOTIFIND are AVG (average of +F and +M neural network scores) and P (probability) scores.
[c]The sensitivity is the percentage of true positives (True+) over the total number of positive patterns in the prediction set (Table I, Column 6). The Specificity is 1 – the percentage of false positives (False+) over the total number of negative patterns in the prediction set (Table I, Columns 7 and 8). Specificity F, Specificity of the full-scale prediction set; Specificity S, Specificity of the small prediction set.
[d]The two highest scoring false positive entries given by each method for the cytochrome c family are: SLTB_BPH30, TA34_TREPA (MOTIFIND); OSM1_YEAST, YEE7_YEAST (BLAST); DHET_ACEPO, ZFH1_DROME (BLIMPS); and COX2_THEP3, COX2_BACSU (HMM).
[e]–, The full-scale prediction was not run.
[f]All false positive entries identified by BLAST and HMM are PETD_XXXX sequences (cytochrome b6-f complex subunit). They lack the PS00192 (cytochrome b_heme) pattern and the heme ligands, but have similarities to the PS00193 (cytochrome b_Qo) pattern. Note that the training sets were compiled based on PS00192, not PS00193.
[g]The specificity of PROSITE is arbitrarily low in these two prediction sets because all false positive patterns identified with PROSITE signatures were included.
[h]The majority of false positive entries identified by BLAST and HMM, and a couple by BLIMPS are NCPR_XXXX (NADPH-cytochrome P450 reductase) and CYSJ_XXXX (sulfite reductase flavoprotein) sequences that have low degrees of sequence similarity to the FMN-binding flavodoxins (PS00201).

distantly related sequence that has a low degree of overall sequence similarity, but a conserved motif region. Examples include CYC4_PSEAE (31.8% identity in 85 aa overlap), CYCL_PARDE (26.5% in 68 aa overlap) and CYB_TRYBB (28.0% identity in 346 aa overlap), all of which have a P score of 0.99 (Table III). MOTIFIND is robust in identifying fragmentary sequences, even those that contain partial or no motif regions, such as CYC_TRYBB, CYB_RABIT, CYB_RANCA, and FLAW_AZOCH. Sequences belonging to under-represented subgroups can also be readily detected, as seen in many cytochrome c entries such as CY2_RHOGE and CY4C_PSEPU.

The only sequences that MOTIFIND is less sensitive in detecting are those of long lengths, usually with > 500 aa, as seen in NIRS_PSEAE and CYB_SULAC (Table III). This problem can be addressed using sliding windows to process query sequences. MOTIFIND works well with small families (i.e. the training set for flavodoxin had only 14 sequences), even with as little as five to ten sequences (results not shown). Furthermore, preliminary studies

Table III. Comparative prediction score results of MOTIFIND and other methods for selected sequence patterns

Sequence ID	Sequence length	ProSite[a]	BLAST	MOTIFIND[b]		BLIMPS	HMM global	HMM motif	Profile[c]
				global/motif	P				
A. Cytochrome C									
Threshold[d]			6.7 E-03	0.43	0.30	287	−5.56	−1.39	4.02
CYC_HORSE	104	T	7.0 E-75	1.00/1.00	0.99	473	144.42	20.83	5.72
CY2_RHOGE	85	T	4.8 E-26	0.26/0.93	0.65	370	0	9.00	4.75
CYC3_DESVM	107	T	3.7 E-74	0.98/1.00	0.99	362	0	6.56	4.37
CYC4_PSEAE	181	T	2.3 E-05	1.00/1.00	0.99	363	2.80	12.44	4.92
CYCP_RHOGE	129	T	6.4 E-28	0.82/1.00	0.97	343	15.74	1.48	4.67
C554_CHLAU	414	T	5.1 E-14	0.01/0.99	0.44	354	0	1.18	4.63
C551_ECTHA	78	T	2.7 E-11	0.61/1.00	0.92	332	−6.46	1.55	4.40
CYCL_PARDE	177	T	2.4 E-04	0.99/0.99	0.99	345	−4.80	6.20	5.22
CY4C_PSEPU	78	T	6.7 E-03	0.72/0.92	0.96	354	−2.09	7.39	4.94
NIRT_PSEST	201	T	8.5 E-83	0.89/1.00	0.99	344	−0.28	9.86	4.73
NIRS_PSEAE	568	T	3.3 E-227	0.01/0.85	0.30	419	7.16	16.62	5.43
YHJA_ECOLI	465	T	4.4 E-61	0.97/0.99	0.99	383	−4.70	4.13	4.80
CYC_CRIFA	113	N	5.0 E-45	1.00/1.00	0.99	339	117.43	10.21	4.27
CYC_EUGVI	102	N	1.2 E-43	0.99/1.00	0.99	295	103.40	2.51	4.01
CY1_EUGGR	243	N	4.2 E-85	0.07/0.99	0.50	287	37.82	3.09	0
C553_PARDE	226	N	3.0 E-18	0.91/0.99	0.97	326	−4.03	−1.39	4.56
CYC_TRYBB	93	P	4.2 E-36	0.99/0.00	0.59	0	85.19	0	0
FDHB_WOLSU	200	F	0	0	0	330	0	0.98	4.26
UROM_HUMAN	640	F	0	0	0	327	0	0	4.40
B. CYTOCHROME B									
Threshold			1.4 E-22	0.43	0.46	948	34.64		11.83
CYB_BOVIN	379	T	6.0 E-268	0.99/1.00	0.99	3373	1097.10	–[e]	26.65
CYB6_MAIZE	215	T	1.3 E-157	0.99/0.99	0.99	2421	370.15	–	18.41
CYB_TRYBB	363	T	1.2 E-55	0.96/0.91	0.99	948	89.08	–	12.26
CYB_PARTE	391	N	1 4 E-22	0.73/0.42	0.86	952	0	–	0
CYB_SULAC	563	N	7.3 E-27	0.47/0.69	0.46	1366	34.64	–	12.26
CYB_RABIT	169	P	1.4 E-106	0.88/0.04	0.77	0	421.19	–	0
CYB_RANCA	114	P	1.4 E-61	0.86/0.01	0.74	0	254.90	–	0
YO43_CAEEL	1040	F	0	0	0	0	0	–	0
C. Flavodoxin									
Threshold			9.0 E-05	0.48	0.89	676	2.65		31.31
FLAV_ANASP	169	T	3.7 E-122	0.99/0.99	0.99	1056	246.47	–[e]	52.25
FLAV_CLOBE	138	T	7.1 E-43	0.93/0.94	0.98	953	69.82	–	32.59
FLAV_MEGEL	137	T	2.2 E-96	0.99/0.99	0.99	955	102.81	–	31.31
FLAW_ECOLI	173	N	1.2 E-61	0.94/0.97	0.97	817	149.24	–	42.89
FLAW_AZOCH	17	P	9.0 E-05	0.00/0.97	0.89	676	2.65	–	0
HEMG_ECOLI	181	F	0	0	0	0	8.64	–	0

[a]The PROSITE codes: (Table I, footnote[b]).
[b]The MOTIFIND scores shown are: 'Global' (+F score, Figure 1), 'Motif' (+M score), and 'P' (probability score). The threshold values shown are AVG (average of +F and +M neural network scores) and P scores.
[c]The PROFILESEARCH results shown are the motif profile results for cytochrome c and b, and global profile results for flavodoxin.
[d]The threshold values shown are the low cut-off scores used in Table II that maximize the sensitivity (detection of true positives) for each given method. Only scores that are greater than or equal to the threshold values are shown; others are given a score of 0.
[e]The prediction was not run.

with several additional protein families, including cytochrome P450, ferredoxin, thioredoxin, G protein receptor, ATPase, and immunoglobulin, all yielded similar predictive accuracies (not shown), indicating that MOTIFIND is generally applicable to other proteins.

The training of the neural networks was fast, ranging from 10 to 50 CPU minutes on the DEC alpha workstation for each family. The training time is proportional to the number of neural interconnections and the number of training sequences. The prediction also ran fast, averaged to ~0.03 CPU second on the workstation for each sequence. The prediction time is dominated by the preprocessing time (i.e. >95% of the total time), and should remain constant regardless of the database size. The time required for preprocessing query sequences is determined by the sequence length and the size of

the n-gram vectors, but independent of the number of training sequences.

Comparative analysis

The accuracy of MOTIFIND is comparable to that of BLAST, but at a significantly faster speed (Table II). On the workstation, the complete SWISSPROT database search by BLAST took between 6 to 10 CPU hours, depending on the number of database sequences (training sequences). But the prediction took < 25 minutes (including preprocessing and postprocessing time) with MOTIFIND, an average speed up of 20 times. MOTIFIND is better than BLAST for identifying short fragmentary sequences containing specific motifs, or distantly related sequences that bear little overall sequence similarity other than the motif regions. The latter is seen in the case for the cytochrome c family.

MOTIFIND is much more sensitive than the PROSITE search, which is based on simple signature patterns to detect family members and runs very fast (Table II). PROSITE search fails to identify motif sequences that are not completely conserved, as defined by the PROSITE signature patterns (i.e. 'N' patterns, Table III); whereas our neural network system is noise tolerant and excellent in handling ambiguous motif patterns. The PROSITE search also fails to detect partial sequences that do not contain specific motifs (i.e. 'P' patterns, Table III); but the detection is possible in MOTIFIND with the incorporation of global information.

The BLIMPS search of BLOCKS also runs fast and is sensitive in detecting family members containing conserved motifs. The method, however, fails to recognize all fragmentary sequences that lack motif regions, including CYC_TRYBB, CYB_RABIT and CYB_RANCA, as one would expect. Furthermore, like the PROSITE search, the number of false positives increases when the BLOCKS/motif length is short, as found in the cytochrome c family. Many false positives returned by PROSITE search ('F' patterns) are also seen in the BLIMPS search result (Table III).

The HMM method can be used for database search of whole proteins or domains. The search using global models is about five to eight times slower than MOTIFIND, as seen in the cytochrome b and flavodoxin families, because it involves the Smith-Waterman search. The global model is less sensitive in detecting distantly related sequences, such as CYB_PARTE (24.5% identity in 351 aa overlap), or fragmentary sequences, such as FLAW_AZOCH (Table III). The global model of cytochrome c also misses several members of the under-represented groups, including CY2_RHOGE, CYC3_DESVM and C554_CHLAU. In contrast, the HMM motif model is more sensitive in detecting the widely diversified members of the cytochrome c family at a much lower false positive rate (Table II). The motif model (with a model length of 15 aa) also runs much faster, because the search time is directly proportional to the model length. The motif model, however, cannot detect fragmentary sequences without motif regions (e.g. CYC_TRYBB), like the BLIMPS search.

The PROFILESEARCH of global profiles is less sensitive, mainly because all short fragments were given low scores, often below cut-off threshold values selected for acceptable specificity. The score can be normalized in relationship to sequence length and used to compute a Z-score if there are more than 400 sequences in the prediction set. However, the use of Z-scores only improves the predictive results marginally, because many sequence fragments become false positives (not shown). Therefore, motif profiles were used in PROFILESEARCH to study both cytochrome c and b families. The results show that it is less sensitive and specific than other motif-based methods, including MOTIFIND, BLIMPS and the HMM motif model (Tables II and III).

Discussion

In this paper, we report a new search method, MOTIFIND, for rapid and sensitive protein family identification. Based on the detailed analysis of false positive and false negative sequence patterns and their predictive scores obtained using various methods, MOTIFIND seems to perform better than other methods for fragmentary sequences, and sequences of under-represented subgroups, many of which are found in the cytochrome c family. For the less diversified cytochrome b family and the conserved flavodoxin family, all methods perform equally well. It should be noted, however, that we did not weigh sequences for PROFILES and BLOCKS heuristically; instead, mostly default parameters were used. Likewise, no extra efforts were made to fine tune the multiple sequence alignments for PROFILES and HMM. Therefore, it is possible that the performance of other methods may be underestimated in the comparative studies. The MOTIFIND, although based on PROSITE patterns derived from alignment, is independent of multiple sequence alignment in generating the neural network model.

The sensitivity of MOTIFIND probably results from the following reasons. First, MOTIFIND uses both positive (member) and negative (non-member) sequences for neural network training in order to enhance sequence discrimination; whereas only positive sequences can be used to build BLOCKS (BLIMPS), models (HMM), or profiles (PROFILESEARCH). Secondly, MOTIFIND uses an integrated neural network design that incorporates both global and motif information as well as long-range

correlation. Furthermore, the n-gram term weight in MOTIFIND is formulated to avoid giving high weights to terms in larger subgroups of a family. Since a large number of protein families in nature are diversified, the sensitivity of MOTIFIND is expected to make it an important tool for family identification.

As a family identification tool, MOTIFIND networks can be easily built for specific families. Due to the small neural network size for each protein family, it is feasible to use MOTIFIND for both on-line training and prediction of any protein families of interest. Both the sequence encoding and neural network designs are general and can be easily extended to improve accuracy. For example, the encoding method can be refined to reflect different motif patterns and to extract long-range correlation of sequence residues by using n-gram terms of different lengths, alphabet sets, distances and their combinations. The neural network can also be expanded to incorporate different sequence discrimination criteria and salient functional/structural patterns. In fact, by combining a sliding window approach and a newly developed encoding method (to be described elsewhere), the MOTIFIND has reached 100% sensitivity at a specificity of 99.95%, 99.99%, and 100%, respectively, for cytochrome c, cytochrome b, and flavodoxin. To fully explore the capabilities of MOTIFIND, we will focus our future studies on the more diversified protein families or domains, such as protein kinase, helicase, and EGF-like domain.

Although still in its early development, MOTIFIND has the potential to become a full-scale DNA/RNA/protein database search and sequence analysis tool. Since more sequences are being generated daily, its speed advantage becomes increasingly significant. In contrast to the database search method that involves pair-wise sequence comparisons and whose search time grows with the number of sequence entries (database size), the search time of MOTIFIND and other family-based search methods only increase with the number of gene families. The current system can be extended into a full-scale protein search tool by adopting the modular neural network design of our protein classification system (Wu *et al.*, 1995). It can also be extended to work on nucleic acid sequences, as demonstrated by our RNA phylogenetic classification system (Wu and Shivakumar, 1994). More studies are needed, however, especially to identify the limits of the protein family size, if there is any. A World Wide Web server will be set up (http://diana.uthct.edu/) to make the MOTIFIND system available to researchers for training and prediction of protein families.

Acknowledgements

The authors thank the referees for their thoughtful suggestions. This study is supported in part by grant number R29 LM05524 from the National Library of Medicine, and by the University Research and Development Grant Program of the Cray Research, Inc. The method described in this paper is the subject of a pending US patent.

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,** 403–410.

Altschul,S.F., Boguski,M.S., Gish,W. and Wotton,J.C. (1994) Issues in searching molecular sequence databases. *Nature Genetics*, **6,** 119–129.

Attwood,T.K., Beck,M.E., Bleasby,A.J. and Parry-Smith,D.J. (1994) PRINTS: a database of protein motif fingerprints. *Nucleic Acids Res.*, **22,** 3590–3596.

Bairoch,A. and Boeckmann,B. (1994) The SWISS-PROT protein sequence data bank: current status. *Nucleic Acids Res.*, **22,** 3578–3580.

Bairoch,A. and Bucher,P. (1994) Prosite: recent developments. *Nucleic Acids Res.*, **22,** 3583–3589.

Doolittle,R.F. (1990) Searching through sequence databases. *Methods Enzymol.*, **183,** 99–110.

Eddy,S.R., Mitchison,G. and Durbin,R. (1995) Maximum Discrimination hidden Markov models of sequence consensus. *J. Comp. Biol.*, **2,** 9–23.

Gribskov,M., McLachlan,A.D. and Eisenberg,D. (1987) Profile analysis: Detection of distantly related proteins. *Proc. Natl Acad. Sci. USA*, **84,** 4355–4358.

Gribskov,M., Luthy,R. and Eisenberg,D. (1989) Profile analysis. *Methods Enzymol.*, **183,** 146–159.

Gribskov,M. and Devereux,J. (1991) *Sequence Analysis Primer*. New York: Stockton Press.

Henikoff,S. and Henikoff,J.G. (1991) Automated assembly of protein blocks for database searching. *Nucleic Acids Res.*, **19,** 6565–6572.

Henikoff,S., Henikoff,J.G., Alford,W.J. and Poetrokovski,S. (1995) Automated construction and graphical presentation of protein blocks from unaligned sequences. *Gene-COMBIS, Gene*, **163,** GC17–26.

Hirst,J.D. and Sternberg,M.J.E. (1992) Prediction of structural and functional features of protein and nucleic acid sequences by artificial neural networks. *Biochemistry*, **31,** 7211–7218.

Karlin,S. and Altschul,S.F. (1993) Applications and statistics for multiple high-scoring segments in molecular sequences. *Proc. Natl Acad. Sci. USA*, **90,** 5873–5877.

Krogh,A., Brown,M., Mian,I.S., Sjolander,K. and Haussler,D. (1994) Hidden Markov models in computational biology: applications to protein modeling. *J. Mol. Biol.*, **235,** 1501–1531.

Pearson,W.R. and Lipman,D.J. (1988) Improved tools for biological sequence comparisons. *Proc. Natl Acad. Sci. USA*, **85,** 2444–2448.

Pearson,W.R. (1991) Searching protein sequence libraries: comparison of the sensitivity and the selectivity of the Smith-Waterman and FASTA algorithms. *Genomics*, **11,** 635–650.

Sarle,W.S. (1994) Neural networks and statistical models. *Proc. 9th Annual SAS Users Group Intn'l Conf.*

Smith,T.F. and Waterman,M.S. (1981) Comparison of bio-sequences. *Adv. Appl. Math.*, **2,** 482–489.

Sonnhammer,E.L.L. and Kahn,D. (1994) Modular arrangement of proteins as inferred from analysis of homology. *Protein Sci.*, **3,** 482–492.

Thompson,J.D., Higgins,D.G. and Gibson,T.J. (1994) CLUSTAL W: improving the sensitivity of progressive multiple sequence alignment through sequence weighting, position specific gap penalties and weight matrix choice. *Nucleic Acids Res.*, **22,** 4673–4680.

Wallace,J.C. and Henikoff,S. (1992) PATMAT: a searching and extraction program for sequence, pattern and block queries and databases. *CABIOS*, **8,** 249–254.

Wu,C.H., Whitson,G., McLarty,J., Ermongkonchai,A. and Chang,T. (1992) Protein classification artificial neural system. *Protein Sci.*, **1,** 667–677.

Wu,C.H. and Shivakumar,S. (1994) Back-propagation and counter-propagation neural networks for phylogenetic classification of ribosomal RNA sequences. *Nucleic Acids Res.*, **22,** 4291–4299.

Wu,C.H., Berry,M., Shivakumar,S. and McLarty,J. (1995) Neural

networks for full-scale protein sequence classification: Sequence encoding with singular value decomposition. *Machine Learning*, **21**, 177–193.

Wu,C.H. (1996) Gene Classification Artificial Neural System. *Methods Enzymol*, **266**, 71–88.

Received on September 7, 1995; revised on December 8, 1995; accepted on December 8, 1995