MOTIF IDENTIFICATION NEURAL DESIGN FOR RAPID AND SENSITIVE PROTEIN FAMILY SEARCH

Cathy H. Wu, Hsi-Lien Chen, Chin-Ju Lo and Jerry W. McLarty

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

Abstract

The accelerated growth of the molecular sequencing data has generated a pressing need for advanced sequence annotation tools. This paper reports a new method, termed MOTIFIND (Motif Identification Neural Design), for rapid and sensitive protein family identification. The method is extended from our previous gene classification artificial neural system and employs two new designs to enhance the detection of distant relationships. These include an n-gram term weighting algorithm for extracting local motif patterns, and integrated neural networks for combining global and local sequence information. The system has been tested with three protein families of electron transferases, namely cytochrome c, cytochrome b and flavodoxin, with a 100% sensitivity and more than 99.6% specificity. The accuracy of MOTIFIND is comparable to the BLAST database search method, but its speed is more than 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 the BLIMPS search of BLOCKS in detecting fragmentary sequences lacking complete motif regions. The method 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 & Devereux, 1991], which may be based on pair-wise comparisons between the query sequence and sequences in the molecular database. These methods range from the most sensitive, but computationally intensive, algorithms of dynamic programming [Needleman & Wunsch, 1970; Smith & Waterman, 1981] to relatively rapid, but less sensitive, methods, such as FASTA [Pearson & Lipman, 1988] and BLAST [Altschul et al., 1990]. 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 & Bucher, 1994], the Profile method [Gribskov et al., 1987], the hidden Markov model [Krogh et al., 1994], and the neural network classification method [Wu et al., 1992; Wu, 1995].

As a database search tool, the family-based (classification) approach has two major advantages over the pair-wise comparison methods [Wu, 1993]: (1) speed, because the search time grows linearly with the number of sequence families, instead of the number of sequence entries; and (2) sensitivity, because the search is based on information of a homologous family, instead of any sequence alone. In addition, the classification approach provides automated family assignment and help organizing second generation databases from which related information can be readily extracted. 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. Some examples include the SBASE protein domain library [Pongor et al., 1994], the BLOCKS database of aligned sequence segments [Henikoff & Henikoff, 1991], the PRINTS database of protein motif fingerprints [Attwood et al., 1994], and the ProDom protein domain database [Sonnhammer & 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 method, termed MOTIFIND (Motif Identification Neural Design), for rapid and sensitive protein family identification, and compare it to the current state-of-the-art methods of the BLAST database search, the PROSITE pattern search, and the BLIMPS search of BLOCKS [Wallace & Henikoff, 1992].

MOTIFIND Design Principals

There are two basic design concepts underlying the new 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

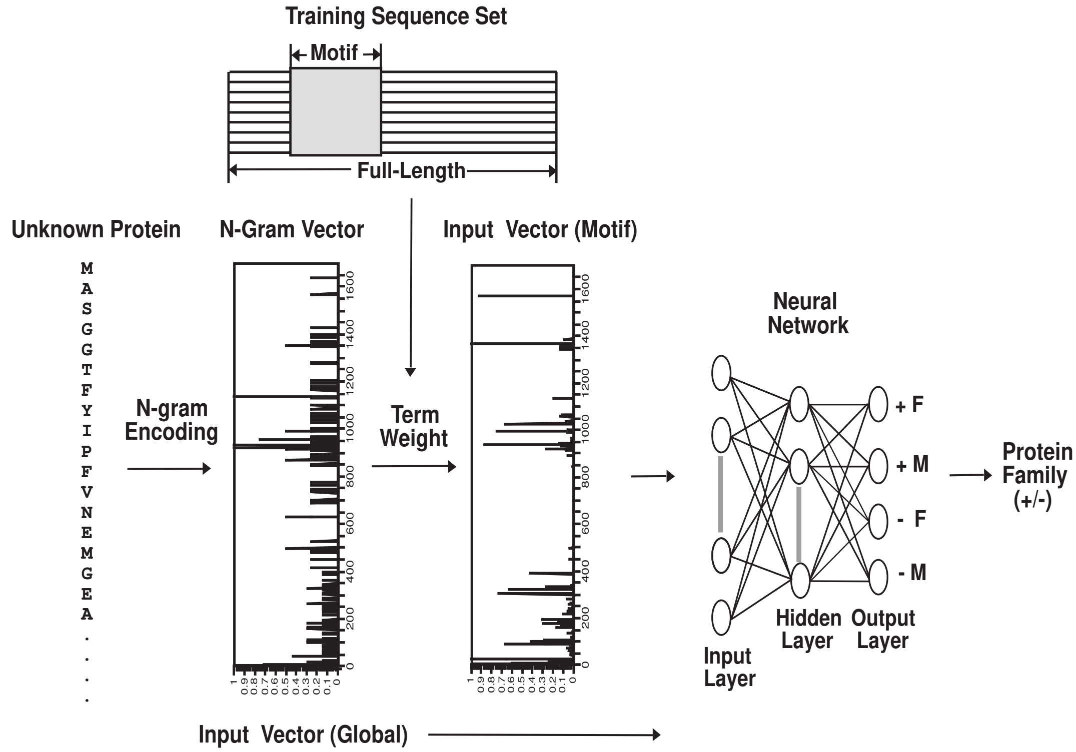

Figure 1. MOTIFIND for rapid and sensitive protein family identification. The sequence strings are converted into input vectors of real numbers (i.e., global and motif vectors) using an n-gram method to encode global sequence similarity and a term weighting method to extract motif information. The neural network then maps the vectors to appropriate output classes according to information embedded in the neural interconnections after network training. Each protein family uses an individual three-layered, feed-forward, back-propagation network.

conserved family information embedded in local 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 introduced two new designs to implement the second concept, an n-gram term weighting algorithm for extracting local motif patterns, and integrated neural networks for combining global (full-length) and local (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, 1993]. Unlike the FastA method, which also uses n-grams (k-tuples), our search method uses the counts, not 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:

$$W_k = \sum_i M_{ik} \; / \; \sum_i F_{ik} \tag{1}$$

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 illustrates 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.

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 & 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. The size of the input layer is determined by the encoding method. 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., $20^2 + 6^4$). Two 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 output layer had four units, representing two parameters (global and motif) for two classes (positive and negative sets). Other network parameters, which were derived from preliminary studies, included: a hidden layer size of 20, random weights of -0.3 to 0.3, a 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. The final neural network architecture was 1696 x 20 x 4, and had 34,000 (i.e., 20 x (1696 + 4)) neural interconnections. Accepted statistical techniques and current trends in neural networks favor minimal architecture (with fewer neurons and interconnections) for its better generalization capability [Le Cun et al. 1990]. Due to the large number of parameters (i.e., weights for the interconnections) to be determined relative to the small number of training patterns for each network (i.e., tens to hundreds), the architecture may not be optimal for generalization. The algorithm we developed for GenCANS to reduce the number of neurons [Wu et al., 1995] will be evaluated for its suitability to MOTIFIND. Meanwhile, because of the large number of input units, instead of using bias units and their additional trained weights, a constant bias term was used.

MOTIFIND Implementation

Program structure. 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.

Data sets. Three protein families of electron transferases, the cytochrome c, cytochrome b and flavodoxin, were used to test the system (Table 1). The positive set consisted of all sequences of the protein family studied, including those 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 & Boeckmann, 1994] by combinations of database sequence search, signature pattern search and manual examination of sequence annotations. The negative set contained all sequences in the SwissProt database that were non-members of the protein family

Table 1. Data sets used for neural network training and prediction.

Protein Family	Prosite Number	Motif Length[1]	Training Set #Positive	#Negative	Prediction Set #Positive	#Negative
Cytochrome C	PS00190	15	149	298	237	43,233
Cytochrome B	PS00192	41	86	172	151	43,319
Flavodoxin	PS00201	19	14	28	23	43,447

[1] 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).

studied. 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 to 2. The ratio was chosen arbitrarily, since the number of negative patterns had little effect on the predictive accuracy as found in preliminary studies where ratios ranging from 1:1 to 1:10 were tested. Approximately two-thirds of the "T" sequences cataloged in PROSITE were chosen randomly as the positive training set ("T" sequences are those containing PROSITE signature patterns). The negative training set were selected randomly from all non-members. The total prediction set is the entire SwissProt database (Release 31.0), containing 43,470 sequences.

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 1) and retrieve substrings in the BLOCKS format [Henikoff & Henikoff, 1991].

MOTIFIND Evaluation Mechanism

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 score and the probability (P) score. The P score was computed using a logistic regression function,

$$\log (P_{hit}/(1-P_{hit})) = \alpha + ß_1 O_1 + ß_2 O_2 + ß_3 O_3 + ß_4 O_4 \tag{2}$$

where P_{hit} is the probability of hit, α, $ß_1$ to $ß_4$ are the regression parameters, and O_1, O_2, O_3 and O_4 are full-length and motif neural network outputs for positive and negative classes, respectively (i.e., +F, +M, -F and -M scores, Figure 1). 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].

A positive sequence is considered to be accurately predicted (i.e., true positive) if both the P score and the average neural network score (i.e., the average of the +F and +M scores) are higher than certain pre-determined threshold values. Conversely, a negative (non-member) sequence is accurately predicted (i.e., true negative) if either score is lower than the threshold. Note that both neural network score and P score range between 0.0 (no match) and 1.0 (perfect match). The SSEARCH program (version 1.7A, July 1994) [Smith & 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 to those obtained by the PROSITE, BLAST and BLIMPS search methods. Different cut-off scores were selected for every method in order to optimize the sensitivity and specificity of each given method. As mentioned above, the prediction set is the entire SwissProt database, containing 43,470 sequences. 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). In PROSITE, the sequences are categorized as "T" (true positive containing signature), "N" (false negative containing degenerated motif not detectable by signature), "P" (false negative lacking motif region, mostly fragmentary), and "F" (false positive containing signature).

The BLAST search was performed using the improved version (version 1.4, October 1994) that adopted Sum statistics [Karlin & Altschul, 1993]. The program was obtained from the NCBI FTP server (ncbi.nlm.nih.gov) and implemented on our DEC alpha workstation running on OSF/1 operating system. The same training set (containing both positive and negative sequences) and prediction set used in MOTIFIND (Table 1) were used as BLAST database and query sequences. The negative set was included as database entries for BLAST search because it provided much better class separation for BLAST (results not shown). The program was run using all default parameters. The result reported was based on the probability score of the first-hit pattern.

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 E-Mail server (blockmaker@howards.fhcrc.org) (version 1.11, June 1994) [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 amino acids (aa), 53 aa and 20 aa, respectively, for the cytochrome c, cytochrome b and flavodoxin families.

Results

Table 2 shows that MOTIFIND achieved 100% sensitivity and more than 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: (1) the degree of overall sequence similarity, (2) the sequence length, (3) the prevalence of the sequence in the family, and (4) the existence of motif region. 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 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. MOTIFIND is robust in identifying fragmentary sequences containing motif regions, such as FLAV_NOSSM (35 aa long, with a P score of 0.99). The method can also find fragmentary sequences 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, CYCP_RHOGE and CY4C_PSEPU.

The accuracy of MOTIFIND is comparable to that of BLAST, but at a significantly faster speed (Table 2). On the workstation, the complete SwissProt

Table 2. Comparisons of the MOTIFIND and other search methods.

Protein Family	Search Method	CPU Time[1]	Sensitivity[2] (%)	True+	Specificity[2] (%)	False+
Cytochrome C	MOTIFIND	984	100.00	237	99.61	167
	BLAST	35,116	100.00	237	99.08	396
	ProSite	27	97.67	231	99.46	233
	BLIMPS	172	99.58	236	98.49	653
Cytochrome B	MOTIFIND	1,452	100.00	151	99.95	23
	BLAST	24,597	100.00	151	99.99	3
	ProSite	33	96.69	146	100.00	1
	BLIMPS	756	98.68	149	99.86	60
Flavodoxin	MOTIFIND	1,019	100.00	23	99.99	5
	BLAST	21,411	100.00	23	99.95	23
	ProSite	34	91.30	21	99.99	5
	BLIMPS	265	100.00	23	99.99	6

[1] The time shown is the total CPU seconds required on a DEC alpha workstation to process the entire prediction set of 43,470 sequences.

[2] The sensitivity is the percentage of true positives (True+) over the total number of positive patterns in the prediction set (Column 6, Table 1). The specificity is 1 - the percentage of false positives (False+) over the total number of negative patterns in the prediction set (Column 7, Table 1).

database search by BLAST took between six to ten CPU hours, depending on the number of database sequences (training sequences). But it took less than 25 minutes (including preprocessing and postprocessing time) on the same machine 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 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 2). PROSITE search fails to identify motif sequences that are not completely conserved, as defined by the PROSITE signature patterns (i.e., "N" patterns); 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); 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 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.

Discussion

In this paper, we report a new search method, MOTIFIND, for rapid and sensitive protein family identification. As a family identification tool, MOTIFIND networks can be easily built and custom-tailored 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, which allows easy expansion and extension. To enhance predictive accuracy, the encoding method can be refined to reflect different motif patterns and to extract long-range correlations of sequence residues by using n-gram terms of different lengths, alphabet sets, distances and their combinations. Furthermore, the neural network can be expanded to incorporate different sequence discrimination criteria and salient functional/structural patterns.

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 & Shivakumar, 1994]. More studies are needed, however, especially to identify the limits of the protein family size, if there is any. A full-scale family identification tool would be especially important to help organize the ever-growing molecular sequence databases according to family relationship.

Acknowledgment

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 U.S. patent.

References

Altschul, S. F., Gish, W., Miler, W.,Myers, E. W. & Lipman, D. J.(1990) Basic local alignment search tool. J. Mol. Biol., 215, 403-410.

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

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

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

Bairoch, A. & Bucher, P. (1994) Prosite: recent developments.Nuc. Acids Res., 22, 3583-3589.

Doolittle, R. F. (1990). Searching through sequence databases. In:Molecular Evolution: Computer Analysis of Proteins and Nucleic Acid Sequences, Methods in Enzymology, Vol 183, R. F.Doolittle, Ed., Academic Press, New York, pp. 99-110.

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

Gribskov, M. & Devereux, J. (eds.) (1991) Sequence Analysis Primer. Stockton Press, New York, 279 p.

Henikoff, S. & Henikoff, J. G. (1991). Automated assembly of protein blocks for database searching. Nuc. Acids Res., 19,6565-6572.

Hirst, J. D. & 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. & Altschul, S. F. (1993) Applications and statistics for multiple highscoring segments in molecular sequences.Proc. Natl. Acad. Sci. USA, 90, 5873 - 5877.

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

Le Cun, Y., Denker, J. & Solla, S. (1990) Optimal brain damage. In Advances in Neural Information Processing Systems 2. San Mateo, CA: Morgan Kaufman.

Needleman, S. B. & Wunsch, C. D. (1970) A general method applicable to the search for similarities in the amino acid sequences of two proteins. J. Mol. Biol., 48, 443-453.

Pearson, W. R. & Lipman, D. J. (1988) Improved tools for biological sequence comparisons. Proc. Nat. 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.

Pongor, S., Hatsagi, Z., Degtyarenko, K., Fabian, P., Skerl, V.,Hegyi, H., Murvai, J. & Bevilacqua, V. (1994) The SBASE protein domain library, release 3.0: a collection of annotated protein sequence segments. Nuc. Acids Res., 22, 3610-3615.

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

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

Sonnhammer, E. L. L. & Kahn, D. (1994) Modular arrangement of proteins as inferred from analysis of homology. Prot. Sci., 3, 482-492.

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

Wu, C. H. (1993) Classification neural networks for rapid sequence annotation and automated database organization. Comp. & Chem., 17, 219-227.

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

Wu, C. H., Berry, M., Shivakumar, S. & McLarty, J. (1995) Neural networks for full-scale protein sequence classification:Sequence encoding with singular value decomposition. Machine Learning, 21, 1-17.

Wu, C. H. (1995) Gene Classification Artificial Neural System. In: Methods In Enzymology: Computer Methods for Macromolecular Sequence Analysis", R. F. Doolittle, Ed., Academic Press, New York, (In Press).