Analysis of DNA Sequence Pattern Using Probabilistic Neural Network Model*

Xiaoming Wu

The Key Laboratory of Biomedical Information Engineering of Ministry of Education
School of Life Science and Technology
Xi'an Jiaotong University, Xi'an 710049, P.R. China
Email: wxm999@hotmail.com
Tel: +86-29-82663454, Fax: +86-29-82660554

Fang Lü, Bo Wang and Jingzhi Cheng

The Key Laboratory of Biomedical Information Engineering of Ministry of Education
School of Life Science and Technology
Xi'an Jiaotong University, Xi'an 710049, P.R. China
Email: {lufang,wbobme,jingzhicheng}@mail.xjtu.edu.cn

To discover frequently occurring DNA patterns related to inherent diseases or gene regulation associated diseases, we must clarify which sequences interact with transcription factors in genome. A probabilistic neural network model was introduced to represent variable length DNA sequence patterns. This model, combined with an EM algorithm, was used to discover conserved sequence patterns from some DNA sequences, and was successfully tested on two datasets, one containing simulated sequences and the other containing upstream sequences of genes in E.coli. Both fixed length and variable length patterns were discovered from the two datasets. The sensitivity of this method was higher than two compared methods, and regulatory sequences of genes were discovered from real DNA sequences of gene clusters. This method could also be used for discovering patterns of protein sequences.

ACM Classification: G.3 (Probability and Statistics-Probabilistic algorithms), I.5.1 (Pattern Recognition – Models – Neural nets), J.3 (Life and Medical Sciences-Biology and genetics)

1. INTRODUCTION

DNA sequences in the human genome comprise many feature elements, which play important roles in gene regulation and the development of many diseases. Through database searches or DNA sequencing, the DNA segment of particular location in the genome can be obtained, accordingly the existence of a special segment relating to a particular disease can be distinguished. However, the sequence pattern, the usual occurring positions, and the biology functions of the element needs to be known previously, which is relatively a difficult task for biologists. Therefore, only a few sequence elements and the corresponding diseases are clearly known. For example, Alu-rich gene *BRCA1*, of which up to 40% of the genomic sequence is composed of Alu element, is involved in

* This work was supported in part by the National Natural Science Foundation of China (No. 60171035)

Manuscript received: 29 July 2004
Communicating Editor: Geof McLachlan

the hereditary predisposition to breast and ovarian cancer; triplet repeat, a type of short tandem repeat sequence, can cause triplet repeat expansion diseases (TREDs) such as Huntington disease and fragile X syndrome (Jasinska and Krzyzosiak, 2004). All cells in the body contain the same set of genes. However, only a fraction of these genes are turned on, or expressed, in an individual cell at any given time. The aberrant expression of certain genes can cause diseases. Once the disease related sequence elements are discovered, they can be identified as targets for medicines to prevent the development of the disease, to remedy the disease, or to alleviate the suffering. New classes of therapeutics can treat a broad spectrum of diseases through affecting the regulation of disease-related genes in patients, or letting enzymes repair or correct the DNA sequence containing disease-related mutations. Some other methods use transcription factor "decoy" strategy as a means of prevention of cardiovascular disease (Morishita *et al*, 1998). All these attempts need to clarify the feature sequences in advance.

The availability of sequences of the human genome and other model organisms allows the comparison of many related DNA sequences simultaneously, which affords a new way to discover those disease related sequences. Computational discovery of the feature sequences is feasible. At present, many methods of discovering fixed length motifs have been introduced. In this paper, we proposed a variable length motif model and used it to discover gapped segments in a given set of DNA sequences.

Methods for computational discovery of motifs vary according to the different kinds of DNA features to be identified, and strategies to discover them. Tandem sequence search, palindrome sequence search, gene identification, functional site discovery and so on, are some examples of them. Hitherto, many discovery methods have been developed. Stormo used a greedy method, combined with a relative entropy measurement to discover universal subsequences in a given DNA set (Stormo and Hartzell, 1989). Liu used a Gibbs sampling search method and employed zero to third-order Markov models to represent background sequences. Liu's method could deal with input data containing noise sequences, and could discover gapped motifs and motifs with palindrome patterns (Liu *et al*, 2001). Another method named AlignACE was based on a similar principle (Hughes *et al*, 2000). By utilizing binding sites features such as sequence similarities and frequencies compared to other sequences, Olman formulated the motif discovery problem as a cluster identification problem, then, designed an algorithm for solving it (Olman *et al*, 2003). Sinha introduced a method which enumerated all motifs in the search space and was guaranteed to produce the motifs with the greatest z-scores (Sinha and Tompa, 2002). Many other methods were introduced in the surveys of Vanet and Pennacchio (Vanet *et al*, 1999; Pennacchio and Rubin, 2001).

Most of the mentioned methods could find fixed length motifs, but in fact, some DNA motifs of variable length also exist, just as protein motifs. For example, Gal4p binding domain has spacers varying in length from 1 to 11bp to bind dimmer or tetramers (Sinha and Tompa, 2002). A candidate transcription factor binding site identified by Pennacchio contains base deletions (Pennacchio and Rubin, 2001). A Hidden Markov Model permits arbitrary gaps in the alignment and is flexible in modeling patterns, but suffers the penalties of added complexity(Lawrence *et al*, 1993). In this article, a probabilistic neural network (PNN) model was introduced to extract fixed length as well as variable length motifs from some given DNA sequences.

2. DESIGN CONSIDERATIONS

The purpose of this article was to provide a novel model and a computer algorithm for the prediction of DNA binding sites, in the condition where a set of sequences were known to contain binding sites for a common factor, but the sites locations were not known. The process was also called "local

multiple alignment" or "pattern discovery" (Stormo, 2000). Some sequences, each of which containing zero to a few instances of a particular sequence pattern, were required to solve the problem. The instances resembled each other, and all of them were similar to an original sequence pattern. In fact, they could be taken as derivations of an original sequence by some manipulations of base replacement or base insertion. In biology, the instances of the sequence pattern usually serve as binding sites of regulatory proteins.

A mathematic model to represent those feature sequences is necessary to make the discovery. Thus, the probability of a subsequence being a feature sequence can be inferred by computing the similarity of the subsequence to the model. Currently, the PWM (Position Weight Matrices) model often serves for the motif model, which uses relative entropy to measure the consistency of feature sequences, but some limitations are inherent in the method (Zhang and Marr, 1993; Benos *et al*, 2002). A good model may comprise more common information of those feature sequences, and can well discriminate ordinary sequences. Novel models are required to make good discoveries. It is also important to use proper similarity measurements to compare the accordance of a candidate sequence to a motif model. The measurement that embodies the true sequence conservation would be a better choice.

Since feature sequences are often scarce in the given sequence set, the search route to find those sequences is equivalently important. In fact, the motif discovery problem has proved to be a NP-hard problem, that is to say, no deterministic polynomial time exists for execution to get the optimal result. An intelligent algorithm can get better results in a relative short period compared to an awkward algorithm. Provided that neither the sequence pattern, nor the total number of the instances and their positions in each sequence are known, the algorithm would guess them initially. The algorithm should distinguish the most frequent element from the input sequences. The discovered element, or the sequence pattern, has a higher chance of being the binding site.

In this paper, a novel motif model was described, and was used to discover frequent occurring, variable length patterns existing in a given set of DNA sequences. Meanwhile, a measurement based on Parzen's probability density estimation, combined with an EM algorithm was used for the discovery.

3. SYSTEM DESCRIPTIONS

3.1 PNN model and algorithm summary

Although many motif models have been used to search feature sequences in DNA, and many successes have been achieved, how to select the right motif model is still a question. PNN is an approach for this classification problem. It overcomes some faults of the traditional back-propagation network (Guo *et al*, 2004). In this study, we applied PNN into the motif discovery field.

In 1990, Spetch proposed the PNN architecture, which was based on Bayesian decision theory and Parzen's method of density estimation (Parzen, 1962; Specht, 1990; Berrar *et al*, 2003). Generally, a PNN is constructed of four layers of neurons (Raghu and Yegnanarayana, 1998), see Figure 1.

The network is made up of four layers. Each layer contains some neural nodes, which are the basic processing units. The input layer is an acceptor, which accepts an unknown vector. The pattern layer contains several neural groups, each corresponding to a recognizable pattern. The weights between the pattern layer nodes and input layer nodes are equivalent to the distance measurements of the two nodes. Each summation layer node sums up the outputs of the corresponding pattern nodes, and a result is made in the output layer node by comparing all the outputs of the summation layer nodes.

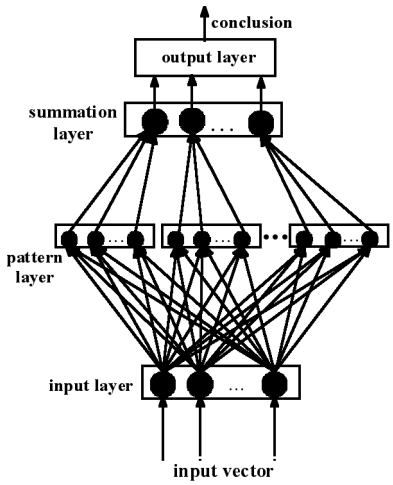


Figure 1: The PNN architecture

Commonly, the input layer contains M nodes and can accept an M-dimensional vector. Each node in this layer is connected to all the nodes in the pattern layer.

The pattern layer comprises K groups of neural nodes and thus makes K patterns. Each node in a pattern corresponds to a sample vector of a known pattern. Every pattern corresponds to a special signal, which can be recognized by the model. This layer also carries out the non-linear transformation of the input vector and the sample vectors. When the PNN model is used to discover DNA sequence patterns, as in this study, vectors in both the input and pattern layers were replaced by different kinds of DNA sequences.

The summation layer has K nodes and can generate K outputs by adding the outputs of all the nodes of each pattern in the pattern layer. Every output of the summation layer nodes correlates to a probability density estimation. According to these probabilities, the output layer makes a decision and gives a result of the input vectors.

The basic method of PNN used to make a discovery is Bayesian inference, which is defined as (Parzen, 1962):

$$P(X \mid f_k) = \frac{1}{(2\pi)^{m/2}\sigma^m \mid f_k \mid} \sum_{X_{ki} \in f_k} \exp[-(X - X_{ki})^T (X - X_{ki}) / 2\sigma^2] \tag{1}$$

This formula calculates the probability of sample X belonging to pattern f_k in the given i sample elements: $X_{k1}, X_{k2}, \ldots X_{ki}$ of that pattern. Here X is the input vector, or the input DNA sequence to be identified. X_{ki} is the i-th sample DNA sequence from pattern k. Originally, m is the dimension of the input vector and pattern vectors; while in this model, it is the average length of the DNA sequence pattern. σ is a smooth factor and $\mid f_k \mid$ is the sample sequence number of pattern f_k.

In Euclidean space, $(X - X_{ki})^T (X - X_{ki})$ is the distance of X and X_{ki}. When the two vectors were used to represent a DNA sequence in this particular model, the other measurement was used to calculate the distance. In (Equation 1), $\exp[-(X - X_{ki})^T (X - X_{ki}) / 2\sigma^2$ is the activate function of the pattern layer nodes. It could also be defined as follows when the nodes were represented as sequences but not vectors:

$$y = \exp[-(d(s_1, s_2))^2 / 2\sigma^2] \tag{2}$$

Here, y is the output of a pattern layer node. s_1 is the sample sequence in motif patterns, s_2 is the input sequence. $d(s_1,s_2)$ is the distance of s_1 and s_2. It could be a Hamming distance when the two sequences are equal in length. In this study, to include variable length sequence pattern, d was the computation of edit distance, which in turn could be calculated by a recursive formula (Vilo, 2002):

$$d(S,\lambda) = d(\lambda,S) = |S|$$
$$d(S,T) = d(S'c,T'b) = \min(d(S',T'b)+1, d(S'c,T')+1, d(S',T')+t(c,b)) \quad (3)$$

Where $S = S'c$, $T = T'b$ and

$$t(c,b) = \begin{cases} 0 & \text{if } c = b \\ 1 & \text{otherwise} \end{cases}$$

All the outputs of the nodes in the same patterns of the pattern layer were summed in the summation layer. As a result, the output of the summation layer nodes would be:

$$y_k = \sum_{s_i \in f_k} \exp[-d(s_1,s_i)^2 / 2\sigma^2] \quad (4)$$

Here, s_1 is the DNA sequence added to the input layer, s_i is a sample sequence represented by a neural node of pattern k. The result y_k is proportional to the probability of the input sequence belonging to the pattern f_k. Only a scale constant $\frac{1}{(2\pi)^{m/2}\sigma^m |f_k|}$ is multiplied.

The smooth factor σ should be selected experimentally. A large value of σ has a better effect on smoothing signals and noises and would eliminate the differences of signals, while a small value would only favour strong signals. In this study, equivalent results have been obtained when σ was selected between 2 and 5.

In this study, two patterns were used in the pattern layer; they were a motif pattern and a background sequence pattern. Nodes in the motif pattern represented short DNA sequences serving as protein binding sites, while the background sequence pattern contained ordinary DNA sequences with no biological functions.

Theoretically the best sample sequences in the background sequence pattern are those picked from genomic DNA sequences, but there needs to be large numbers of nodes in the pattern layer, and it requires massive amounts of computer time. As an alternate, a zero-order HMM model, whose parameters were derived from realistic genomic sequences, was used to generate the sample sequences.

The outcome was produced in the decision layer by considering the two probabilities of outputs from the two summation layer nodes. Final results were obtained. They reflected the character of the sequence: a conserved sequence or an ordinary sequence.

3.2 Experimental steps

Similar to other motif discovery methods, a set of DNA sequences, each of which contained some protein binding sites, was used as input data. The difference was that the length of the binding sites could vary. So the alignment of the discovered sequences would contain gaps. Four major steps of an EM algorithm were used for the discovery.

1) Model initialization

According to the average length of the sequence pattern to be identified, the number of nodes of the input layer was determined in the first place. At the same time, a given number of sample sequences

representing a motif pattern were randomly selected from the input sequences by the algorithm. These sample sequences were used to represent the common characteristics of the feature sequence, the counts of which were equal to the total number of potential binding sites in the input sequences. Sequence fragments in the background pattern were randomly picked from artificial genomic sequences generated by a zero-order HMM model. The length of each sequence fragment and the total number of sequences were determined according to parameters given by the user. By now, the model included two patterns: the motif pattern and the background sequence pattern. The sample sequences in the motif model were not conserved, and were less effective in representing any sequence pattern. Consequently, other refining steps were needed to optimize the motif pattern.

2) Candidate sequences selection (Expectation step)

By comparing all the subsequences in the dataset with the current pattern, we could evaluate the identity of each subsequence to the motif pattern. The ones whose length were close to the motif, were selected and placed on the input layer of the PNN. Particularly in this study, if the desired motif length was l, all the subsequences at the length of l–1,l, l+1 in the dataset would be selected. Then, the scores of each subsequence to the motif pattern and the background sequence pattern were calculated by the summation layer nodes. The scores were the expectation values of the input sequence being either the motif pattern or the background sequence pattern. A judgment was then made in the decision layer according to the two scores, and a conclusion that the subsequence is either a potential binding site or a background sequence would be produced. Those identified as potential binding sites were selected and used to update the current pattern. This step includes computing expectations of subsequences being the sequence pattern, so it could be taken as the expectation step.

3) Pattern update (maximization step)

The new identified subsequences in the previous step were more identical and more conserved than the sample sequences in the current motif pattern, and would represent a better sequence pattern. Replacing all the sample sequences in the motif pattern with the new identified ones, the expectation of subsequences being the sequence pattern would be maximized and the sequence pattern revised. This step is also the maximization step.

4) Stop condition

After step 2 and step 3 were repeated many times, more conserved subsequences would be discovered. When the cycle reached a given number, or the discovered subsequences converged, the process would be finished automatically, and a conserved sequence alignment representing a binding site would be obtained.

The procedure underwent many Expectation-Maximization steps of the EM algorithm, and could converge to a local optimal result. By using different initial data and running the algorithm many times, results close to the global optimal would be obtained. An Excel add-in named PNNMD implementing the algorithm was developed and used for tests. Users could load sequences and make the discovery from a sample interface. The Excel Add-in with some example data can be obtained from the location: http://202.117.57.48/pnnmd/, or by sending a request by email to the authors.

4. RESULTS

Experiments with simulated data

"Challenge problem" is representative in motif discovery. The object is to discover all the fixed length subsequences, each of them comprised of some base mismatches to an original sequence in

the given dataset (Pevzner and Sze, 2000). Here, we defined an alternative way to describe an unfixed length motif existing in the dataset.

Definition: In *N* pieces of DNA sequences, a total of *M* unfixed length motif instances exist, they all originate from *K* edit manipulations of a same DNA sequence of the length of *L*. Each edit manipulation is a base deletion or replacement. This kind of motif could be written as *M (L, K)*. For example: after two edit manipulations, the sequence *tagcat* would produce motif instances: *tGgcAt, tC_cat, tag_Ct, tCGcat*. Here, a capital letter means base replacement and "_" means base deletion.

We generated two sequence sets, each of them containing 30 motif instances. The motifs in the two sets were *M (10,2)* and *M (12,3)*, respectively. Other bases in the sequence sets were generated from an equal probability independently identical distribution (iid). That is, 4 bases A, C, G and T were equally generated. One dataset contained 20 sequences of the length of 200bp; the other contained 30 sequences of the length of 300bp. During the datasets construction, the motif instances were randomly inserted in the background sequences. Therefore, some sequences could hold multiple instances, while others would hold no instance, which was close to a realistic situation.

Three different motif lengths 9, 10 and 12 were used to make the discovery on the two datasets. The initial sample sequences of motif patterns in each discovery were randomly selected by the algorithm from the input dataset. The number of motif instances, or the nodes, in the motif pattern were set from 20 to 30. The sample sequences number in the background sequence pattern was set to 200, and the length of those sequences was equal to the expected motif length.

The combination of different motif lengths and numbers of motif instances was used to construct parameter sets. The discovery was performed 100 times by using each parameter set, and the best results are selected and listed in Table 1.

Furthermore, we used two representative methods MEME (Bailey and Elkan, 1995) and AlignACE to make the same discovery as comparisons. The parameters and results are also provided in Table 1.

In Table 1, *L* is the motif length; and *nSites* is the motif instances in the dataset. True positive (tp) means those whose locations are correctly recognized, false positive (fp) is those whose locations are wrongly recognized. False negative (fn) is true motifs, which are not recognized by the methods. The sensitivity of the algorithm is calculated by: $Sn = tp/(tp + fn)$.

From Table 1 we can see: In dataset 1, the best sensitivity by this method is 0.83, higher than MEME and AlignACE method, which are only 0.77 and 0.8. Dataset 2 comprises longer motifs and longer input sequences, while the *nSites* is the same. At this time, the best sensitivity is 0.93, also better than the results of other two methods. It is clear that if the motif length and motif number in the parameters are close to the true situation, better results are likely to be obtained. Since edit

Datasets		PNN												MEME				AlignACE			
1	seq.No:20 nsite:30 length= 200bp	*L=9,nSites=20*				*L=10,nSites=25*				*L=12,nSites=30*				*minw=8,maxw=13 minsites=20 ,maxsites=30*				*numcols=10 expects =30, gcback=0.6*			
		tp	fp	fn	sn	tp	fp	fn	sn	tp	fp	fn	sn	tp	fp	fn	sn	tp	fp	fn	sn
		18	2	12	0.60	24	1	6	0.80	25	5	5	0.83	23	4	7	0.77	24	6	6	0.80
2	seq.No:30 nsite:30 length= 300bp	*L=9,nSites=20*				*L=10,nSites=30*				*L=12,nSites=30*				*minw=10, maxw=14 minsites=25, maxsites=35*				*numcols=10 expects =30, gcback=0.6*			
		tp	fp	fn	sn	tp	fp	fn	sn	tp	fp	fn	sn	tp	fp	fn	sn	tp	fp	fn	sn
		15	5	15	0.50	28	2	2	0.93	26	4	4	0.87	25	3	5	0.83	26	4	4	0.87

Table 1: Comparison of PNN with MEME and AlignACE using simulated data

distance was used in the algorithm, some motifs with 1 to 2 gaps were discovered, while the other two methods ignored that kind of motif. This is likely to be one reason for the improvement.

Experiments with true DNA sequences

In the realistic DNA, the sequence form is much more complex, and the discovery becomes more difficult. One of the important applications of motif discovery is to identify conserved sequences in the upstream regions of sets of co-expressed genes obtained from microarray data cluster. Spellman analyzed some microarray data of yeast cell-cycle related experiments, and identified many gene clusters. In them, the *CLN2* cluster is a large cluster comprising many genes that are subject mainly to cell cycle regulation. These genes were considered to have common binding sites to some proteins in the regulatory regions (Spellman *et al*, 1998). We picked up all upstream regions from -1 to -700bp relative to transcription start sites of a total of 57 genes, and constructed a dataset, then used this method to make the discovery.

In the discovery, the motif length was set to from 6 to 10, and the nodes in the motif pattern were set from 20 to 50, representing 20–50 motif instances. The background sample sequences came from synthetic genomic data that was generated from a zero-order HMM. The sequences number was set to 100, and the sequences length was equal to the expected motif length. Each discovery began with different initial sample sequences, and the procedure was repeated 100 times, some of the best results were selected and arranged in Table 2.

In Table 2, five discovered motifs and the genes that comprise them are listed. The alignments came from some representative sample sequences of the final motif pattern. Spellman pointed out that three important motifs exist in the dataset. They were motif ACGCGT, which was the MCB element, motif CRCGAAA, which was the SCB element, and AGAAGAAA, which was a conserved sequence with unclear function. From Table 2 we can see, both SCB and MCB elements

motif	Parameters	Alignments	Genes	Functions
1	*l=6 nsites=30*	TCGCGT ACGTGT ACGCGC ACGCCT ACG.CT	UNG1, TOF1, SRO4, SPT21, SPO16, SPH1, SMC3, SMC1, RNR1, RHC18, RFA2, RFA1	MCB elements
2	*l=6 nsites=25*	TCGAAA T.GAAA GCGAAA CCGA.A ACGAAA	TOF1, SWE1, STB1, SRO4, SPT21, SPO16, SPH1, RSR1, RNH35, RFA2, RFA1, POL30, CLN2, CDC9	SCB elements
3	*l=8 nsites=20*	TTCCTTTT TCCCTTTT GTCCTTGC TCCCTTGA	YOX1, RFA2, RFA1, RAD53, RAD27, POL30, MSH6, CAC2, ASF2	Unknown, DB entries: r04339, r03553
4	*l=8 nsites=50*	TATATAGA TATATATA TATATAAA	POL30, MCD1, RNR1, KIM2, HIF1, BNI4, TOF1, MSH2, CLN2	TATA box
5	*l=9 nsites=20*	GGTTTTATG GCTTTTAAG CCTTTTAAA GCTTTTAAA	CDC9, CLB6, SWE1, HIF1, CDC45, RAD53	Unknown, DB entries: r01598

Table 2: Feature sequences discovered from CLN2 gene cluster

In this table, *l* is the motif length; and *nsites* is motif instances in the dataset. DB entries in the last column refer to TRANSFAC entries.

were discovered by this method, together with a TATA box element. In addition, two conserved sequences with unknown function were discovered. One was motif 3, the consensus was KTCCTTKT; the other was motif 5, the consensus was SCTTTTAAR. Both occurred many times in the dataset. Through TRANSFAC database searching, we matched the two motifs to 3 TRANSFAC entries, which meant that in other species, these kinds of sequences could arouse gene regulation. With some confidence, we can guess that, the two conserved sequences would have an analogous function in yeast.

This algorithm was also tested by using other gene cluster datasets. Many conserved sequences were discovered, and lots of them could match TRRD and TRANSFAC databases, indicating the protein binding property of those sequences. As a conclusion, this method can be considered as a useful tool for searching for potential functional segments in DNA sequences.

5. CONCLUSION

We present a sequence pattern discovery method based on PNN and EM algorithm to discover variable length conserved segments hidden in DNA sequences. The method shows some advantages when compared with other representative methods such as MEME and AlignACE in tests of two simulated datasets. In true DNA sequence analysis, motifs coinciding with reality were also discovered, and two more conserved sequences were identified, which could match TRANSFAC entries and be inferred to have some potential biological functions.

A probability value was used to evaluate the likelihood of a subsequence belonging to a sequence pattern, which was a natural statistical inference of the sequence. Instead of sequence models, real DNA subsequences are stored in the network. Consequently, all features of those sequences are preserved. Since a matrix based model can be used only for searching fixed length motifs, this model shows an improvement and can be used to search flexible length motifs in a genome.

This method could also be regarded as a framework of a new way to explore common signals in biological sequences. Specially, edit distance used in this method could be replaced by other measurements such as Levenshtein distance. As a result, the cost of base substitution, base insertion or base deletion could be set to different values, which is rational in some circumstances. More complex similarity measurements, which would embody the features of inter-base relationships, or base order occurrence chance of the candidate subsequences, could also be used. In addition, the search route, as an EM algorithm used in this study, could be replaced by other intelligent algorithms to achieve better search results. Through using different background sequence patterns and adding more sequence patterns, subtle signals which might be miss-recognized by other methods would be discovered.

Since the PNN model is based on statistics and probability inference, it can be extended to solve different kinds of problems. The future work is to apply this model to protein domain discovery and RNA structure prediction.

REFERENCES

BAILEY, T. L. and ELKAN, C.(1995): Unsupervised learning of multiple motifs in biopolymers using expectation maximization, *Machine Learning*, 21(1–2):51–80.

BENOS, P. V., BULYK, M. L. and STORMO, G. D.(2002): Additivity in protein-DNA interactions: how good an approximation is it? *Nucleic Acids Research*, 30(20):4442–4451.

BERRAR, D. P., DOWNES, C. S. and DUBITZKY, W. (2003): Multiclass cancer classification using gene expression profiling and probabilistic neural networks, In *Pacific Symposium on Biocomputing*. 8:5–16.

GUO, J., LIN, Y. and SUN, Z. (2004): A novel method for protein subcellular localization based on boosting and probabilistic neural network, In *Proc. Second Asia-Pacific Bioinformatics Conference* (APBC2004), Dunedin, New Zealand. CRPIT, 29. Chen, Y.-P. P., Ed. ACS. 21–27.

HUGHES, J. D., ESTEP, P. W., TAVAZOIE, S. and CHURCH, G. M.(2000): Computational identification of cis-regulatory elements associated with groups of functionally related genes in saccharomyces cerevisiae, *Journal of Molecular Biology*, 296(5):1205–1214.

JASINSKA, A. and KRZYZOSIAK, W. J.(2004): Repetive sequences that shape the human transcribtome, *FEBS letters*, 567:136–141

LAWRENCE, C. E., ALTSCHUL, S. F., BOGUSKI, M. S., LIU, J. S., NEUWALD, A. F. and WOOTTON, J. C.(1993): Detecting subtle sequence signals: a Gibbs sampling strategy for multiple alignment, *Science*, 262:208–214.

LIU, X., BRUTLAG, D. L. and LIU, J. S. (2001): BioProspector: discovering conserved DNA motifs in upstream regulatory regions of co-expressed genes. In *Pacific Symposium on Biocomputing*. 6:127–38.

MORISHITA, R., HIGAKI, J., TOMITA, N. and OGIHARA, T.(1998): Application of transcription factor "decoy" strategy as means of gene therapy and study of gene expression in cardiovascular disease, *Circulation Research*, 82(10):1023–1028.

OLMAN, V., XU, D. and XU, Y.(2003): CUBIC:Identification of regulatory binding sites through data clustering, *Journal of Bioinformatics and Computational Biology*, 1(1):21–40.

PARZEN, E.(1962): On estimtion of a probability density functiuon and mode, *Ann Math. Statist.*, 33(6):1065–1076.

PENNACCHIO, L. A. and RUBIN, E. M.(2001): Genomic strategies to identify mammalian regulatory sequences, *Nature Reviews Genetics*, 2(2):100–109.

PEVZNER, P. A. and SZE, S. H. (2000): Combinatorial approaches to finding subtle signals in DNA sequences, In *Proceedings of 8th International Conference on Intelligent Systems for Molecular Biology (ISMB)*. 269–278.

RAGHU, P. P. and YEGNANARAYANA, B.(1998): Supervised texture classification using a probabilistic neural network and constraint satisfaction model, *IEEE Transactions on Neural Networks*, 9(3):516–522.

SINHA, S. and TOMPA, M.(2002): Discovery of novel transcription factor binding sites by statistical overrepresentation, *Nucleic Acids Research*, 30(24):5549–5560.

SPECHT, D. F.(1990): Probabilistic neural networks, *Neural Networks*, 3:109–118.

SPELLMAN, P. T., SHERLOCK, G. and ZHANG, M. Q.(1998): Comprehensive identification of cell cycle-regulated genes of the yeast saccharomyces cerevisiae by microarray hybridization, *Molecular Biology of the Cell*, 9(12):3273–3297.

STORMO, G. D. (2000): DNA binding sites: representation and discovery, *Bioinformatics*, 16(1):16–23.

STORMO, G. D. and HARTZELL, G. W.(1989): Identifying protein-binding sites from unaligned DNA fragments, *Proceedings of the National Academy of Sciences of the United States of America*, 86:1183–1187.

VANET, A., MARSAN, L. and SAGOT, M. F.(1999): Promoter sequences and algorithmical methods for identifying them, *Research in Microbiology*, 150:779–799.

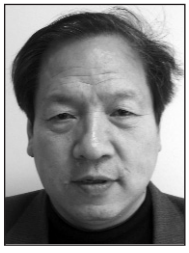
VILO, J. (2002): Pattern discovery from biosequences, Department of Computer Science, University of Helsinki. Finland, University of Helsinki.

ZHANG, M. Q. and MARR, T. G.(1993): A weight array method for splicing signal analysis, *Computer Applications in the Biosciences*., 9(5):499–509.

BIOGRAPHICAL NOTES

Xiaoming Wu

Xiaoming Wu is a PhD candidate and a research assistant at the Department of Biomedical Engineering, Xi'an Jiaotong University, P.R.China. He received his MSc degree in Biomedical Engineering from Jiaotong University in 2000. His research interests are medical image processing, medical image transmission and biosequences analysis. He has published several articles in journals and several papers at international conferences.



Bo Wang

Bo Wang is a professor and PhD supervisor in the Biomedical Engineering Department, Xi'an Jiaotong University, P.R.China. His research interests include ultrasound signal processing, ultrasound medical diagnosis, ultrasonography, image feature extraction. Currently, he is the director of the Research Center of Bioinformatics and deputy director of the Department of Biomedical Engineering. He is also a senior member of the Chinese Institute of Biomedical Engineering (CIBE) and has published 40 papers in journals and proceedings of international conferences.

Fang Lü is an associate professor in the Biomedical Engineering Department, School of Life Science and Technology, Xi'an Jiaotong University, P.R.China. She received her MD in Xi'an Medical University, China, in 1982. She also worked as a visiting scientist at the University of Oklahoma, USA, from 2002 to 2003. Her research interests include biologic signal processing and the mechanism of pain modulation. She is a member of American Society for Neurosciences. She is involved in several research projects as responsible and senior researcher related to these subjects.

Fang Lü

Jingzhi Cheng is a professor in the Department of Biomedical Engineering, School of Life Science and Technology, Xi'an Jiaotong University, P.R.China. He is one of the founders and group leaders of Biomedical Engineering discipline in China. His current research interests are applications of generalized time-serial analysis on biomedical signal processing and ultrasonic signal processing. He has served on the editorial board of some domestic Chinese journals.

Jingzhi Cheng