Finding Genes by Hidden Markov Models with a Protein Motif Dictionary

Kiyoshi Asai [1] asai@etl.go.jp

Tetsushi Yada [2] yada@tokyo.jst-c.go.jp

Katunobu Itou [1] kito@etl.go.jp

[1] Genome Informatics Group, Electrotechnical Laboratories (ETL)
1-1-4 Umezono, Tsukuba, Ibaraki 305 Japan

[2] Japan Science and Technology Corporation(JST)
5-3 Yonbancho, Chiyoda-ku, Tokyo 102 Japan

Abstract

A new method for combining protein motif dictionary to gene finding system is proposed. The system consists of Hidden Markov Models (HMMs) and a dictionary. The HMMs represents the nucleotide acid bases, the codons, and the amino acids. The 'words' in the dictionary is described by the sequence of these HMMs and represent the non-coding regions, the codons, protein motifs, tRNA regions and signals in DNA sequences. The statistics between these regions are expressed by the "grammar", which is a stochastic network of the 'words.'

Using the same kind of technique of speech recognition by HMMs with a word dictionary and a grammar, the stochastic network of 'words' enables the motif dictionary to be used during the parsing of the DNA sequences. At the same time, the information of the di-codon statistics, which are known as the important parameters, is included in the stochastic network. As a result, while the system parses DNA sequences and finds the coding regions, the protein motifs are automatically annotated in the regions. It helps to identify the functions of the genes and reduces the cost of homology search for each hypothetical coding regions. This method is different from simply using the the information of homology search. This method uses the information of the motif patterns during *the parsing process, but searching the motif patterns* after/before *finding the coding regions cannot directly affect the parsing process itself. Experimental results have shown that this method correctly finds and annotates the motifs in the coding regions in the DNA sequence of cyanobacterium.*

1 Introduction

The progress of the sequencing projects and the resulting large sequence data demands the computational biologists to develop effective systems to detect genes in the DNA sequences. ([2]). In this paper, we propose a new method for combining a protein motif dictionary to gene finding system based on hidden Markov models (HMMs).

1.1 HMMs in gene findings

Because genes have a structure like a language, linguistic methods are effective ([3]). However, the components and the rules of the 'DNA language' are non-deterministic, it is necessary to combine the statistics and the linguistics for the 'parsing' of DNA. That is why hidden Markov models (HMM) are becoming widely used for gene recognition ([9][10][13][14]).

In order to build a stochastic "DNA language" by using HMMs, we model the components of the gene structure by HMMs and connect these HMMs by the rules which represents the gene structure. From a view point of stochastic grammar, a HMM is a stochastic regular grammar. Regular grammar can be expressed by the networks of the symbols. A nice feature of regular grammar is its modularity. A network of the networks which represent regular grammars becomes a regular grammar. HMMs have the same property: a network of the networks which represent HMMs becomes an HMM. If we model the promoters, codons, amino acids, motifs and other objects on DNA by HMMs (for example, Figure 3), the networks of these objects form a new HMM (for example, Figure 1). This means we can parse the whole DNA sequence by the combined HMMs using a dynamic programming algorithm (Baum-Welch algorithm). Same kind of parsing was used for protein structure prediction in [1].

1.2 Previous work

One of the authors developed a gene finding system based on hidden Markov models to detect the protein coding regions within 1M base continuous sequence data of cyanobacterium, *Synechocystis* sp. strain PCC6803, and achieved the recognition accuracy 90.7% for coding regions and 88.1% for intergenic regions ([14]). The recognition performance of that work was good, although the implementation of the models was simple. In that system, the parameters of each HMMs had been decided separately, using the concept of modularity described in Section 1.1. That is analogous to the training of parameters 'with labels' in speech recognition [15]. The main statistics between HMMs were bigrams, which is a first order Markov model (not 'hidden' Markov model).

1.3 HMMs with a motif dictionary

In order to build an effective gene recognition system, it is important to build good models of genes. Many parameters for the modeling of coding/non-coding regions have been proposed, most of which represent the local characteristics of the genes. However, it is more desirable to have stochastic models of these proteins than to have merely the local statistics of the genes, because the coding regions are translated into proteins and the sequences of coding regions have the feature of the real amino acid sequences of proteins. The significant improvement taken in this paper is that we have combined stochastic protein motif models to the gene recognition system, while the HMM gene models are basically same as used in the previous work [14].

By using a motif dictionary as a component of the system, the motif names are annotated on the candidate of the coding regions (Figure 2). Note that to have protein models with the system is different from the popular technique of homology search of the hypothetical coding regions. The latter searches the database *before or after* the system decides the candidate of the coding regions, while the former uses the information of the database *during* the process of deciding the candidate of the coding regions.

We have built the gene recognition system using speech recognition software (HTK[15]). The codons, amino acids, intergenic models are built by HMMs, taking the natural output symbols as the four kinds of bases, 'A','C','G','T'. Using these HMMs as 'phonemes,' we have constructed the 'word' dictionary, whose vocabulary is the protein motifs. The recognition process is exactly the same as the dynamic programming parsing of the speech, using a grammar defined on these 'words.'

2 Data

2.1 DNA sequences

We used the same DNA sequence as that of the previous work ([14]), a contiguous sequence of 1M base of a unicellular cyanobacterium, *Synechocystis* sp. strain PCC6803 for the gene recognition. The sequence is divided into 8 entries in GenBank (D63999 - D64006) with potential protein coding regions in the annotations ([8][7]). We respected the eight divisions of the sequence and used each division as a test set. Training data sets are created from the remaining 7 divisions, excluding the division of test set. The data of training sets determine the HMM parameters and a test set was used to validate the recognition ability of HMM based on parameters derived from the training set. The parameters include those for the coding regions and those for the intergenic regions by using the annotations of the coding regions in the entries. Each entry was used twice as a test set: a test set to evaluate detection of coding in the normal direction and that to evaluate detection in its complementary direction.

2.2 Protein motifs

In order to construct a motif dictionary for the gene recognition system, we extracted 1149 motif entries from PROSITE release 13.0([11]), and selected 933 motif patterns as the candidate of the 'words' in the motif dictionary. We selected these patterns according to an evaluation score based on the specificity of the patterns. For example, `A-[PN]-S-[VIL]` is 20/1 specific ('A' and 'S') in two positions, 20/2 ('[PN]') and 20/3 ('[VIL]') specific in each position. The overall specificity is the product of these values. The higher specificity is preferable in order to avoid the 'pattern match by chance.' We searched the data of annotated coding regions described in Section 2.1 by these 933 motif patterns, and found 156 hits of 76 patterns. We adopted these 76 patterns (Table 1) as the vocabulary of the motif dictionary for the gene recognition system.

Table 1: Vocabulary of the motif dictionary

PS00021	KRINGLE	PS00028	ZINC FINGER C2H2	PS00029	LEUCINE ZIPPER
PS00039	DEAD ATP HELICASE	PS00054	RIBOSOMAL S11	PS00074	GLFV DEHYDROGENASE
PS00093	N4 MTASE	PS00097	CARBAMOYLTRANSFERASE	PS00138	SUBTILASE SER
PS00154	ATPASE E1 E2	PS00157	RUBISCO LARGE	PS00163	FUMARATE LYASES
PS00168	TRP SYNTHASE BETA	PS00171	TIM	PS00179	AA TRNA LIGASE II 1
PS00188	BIOTIN	PS00196	COPPER BLUE	PS00198	4FE4S FERREDOXIN
PS00244	REACTION CENTER	PS00296	CHAPERONINS CPN60	PS00297	HSP70 1
PS00329	HSP70 2	PS00337	BETA LACTAMASE D	PS00343	GRAM POS ANCHORING
PS00365	NIR SIR	PS00370	PEP ENZYMES PHOS SITE	PS00381	CLP PROTEASE SER
PS00382	CLP PROTEASE HIS	PS00395	ALANINE RACEMASE	PS00442	GATASE TYPE I
PS00461	6PGD	PS00480	CITRATE SYNTHASE	PS00534	FERROCHELATASE
PS00571	AMIDASES	PS00600	AA TRANSFER CLASS 3	PS00614	IGPS
PS00632	RIBOSOMAL S4	PS00665	DHDPS 1	PS00666	DHDPS 2
PS00667	COMPLEX1 ND1 1	PS00668	COMPLEX1 ND1 2	PS00674	AAA
PS00693	LUM BINDING	PS00698	GLYCOSYL HYDROL F9 2	PS00704	PROK CO2 ANHYDRASE 1
PS00715	SIGMA70 1	PS00742	PEP ENZYMES 2	PS00806	ALDOLASE CLASS II 2
PS00815	AIPM HOMOCIT SYNTH 1	PS00816	AIPM HOMOCIT SYNTH 2	PS00839	SUMT 1
PS00840	SUMT 2	PS00844	DALA DALA LIGASE 2	PS00846	HTH ARSR FAMILY
PS00859	GTP CYCLOHYDROL 1 1	PS00860	GTP CYCLOHYDROL 1 2	PS00866	CPSASE 1
PS00870	CLPAB 1	PS00871	CLPAB 2	PS00889	CNMP BINDING 2
PS00893	MUTT	PS00906	UROD 1	PS00921	NITRIL CHT 2
PS00936	RIBBOSOMAL L35	PS00937	RIBBOSOMAL L20	PS00954	IGP DEHYDRATASE 1
PS00955	IGP DEHYDRATASE 2	PS01042	HOMOSER DHGENASE	PS01065	ETF BETA
PS01066	YBR002C	PS01071	GRPE	PS01094	YER057C YJGF
PS01096	PPIC PPIASE	PS01118	SUI1	PS01136	YHDG
PS01139	BACT MICROCOMP				

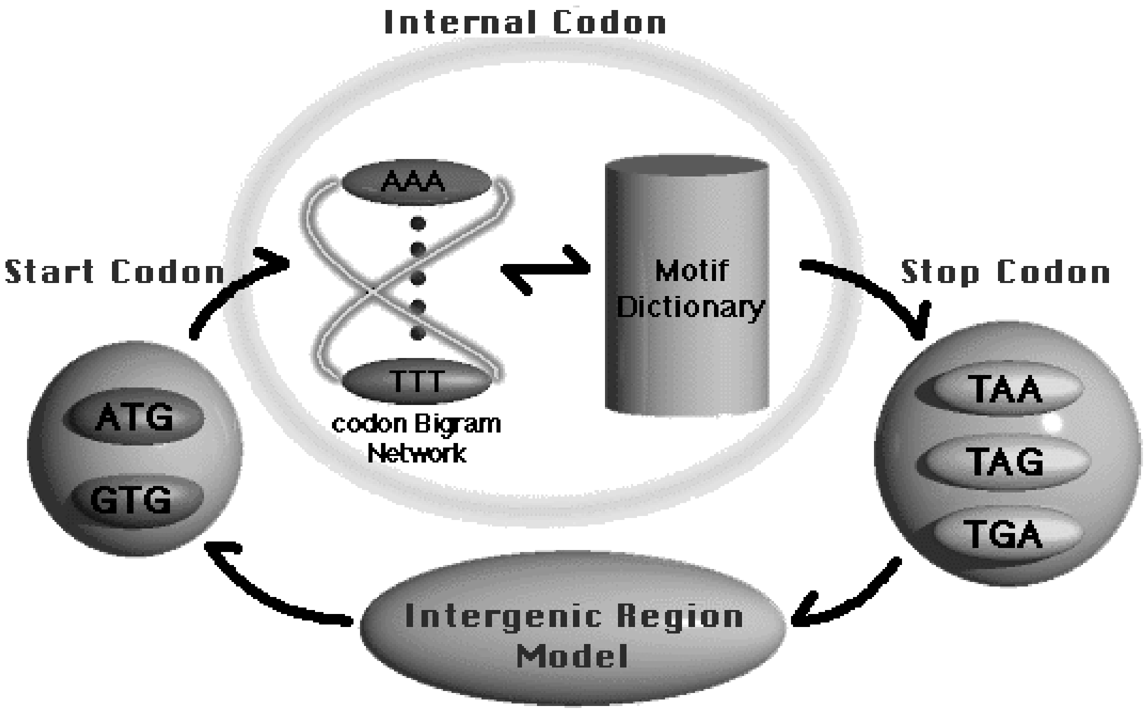


Figure 1: Overview of the gene recognition system

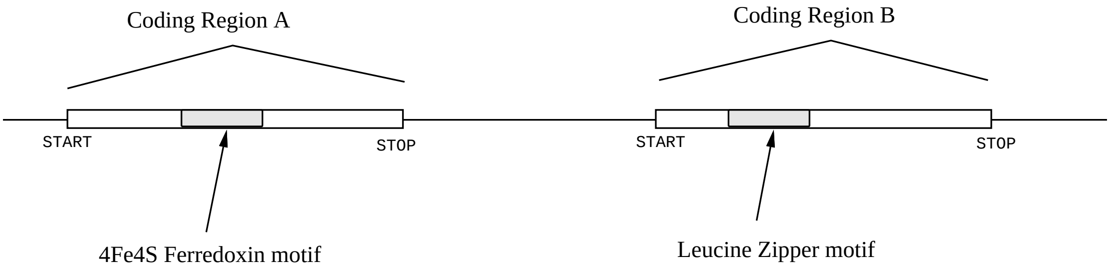
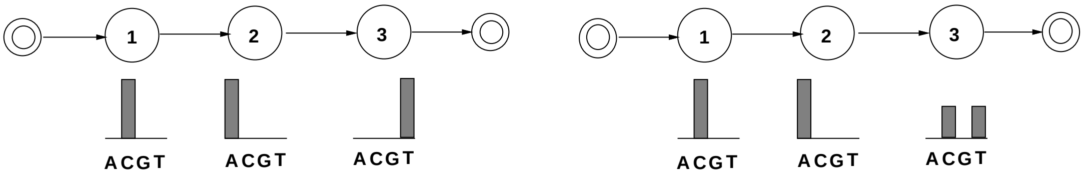


Figure 2: Sketch of gene recognition with motif annotation



(a) HMM of codon CAT

(b) HMM of amino acid Histidine

Figure 3: Examples of HMMs of a codon and an amino acid. The states expressed by the double circles are special state of null output.

3 System

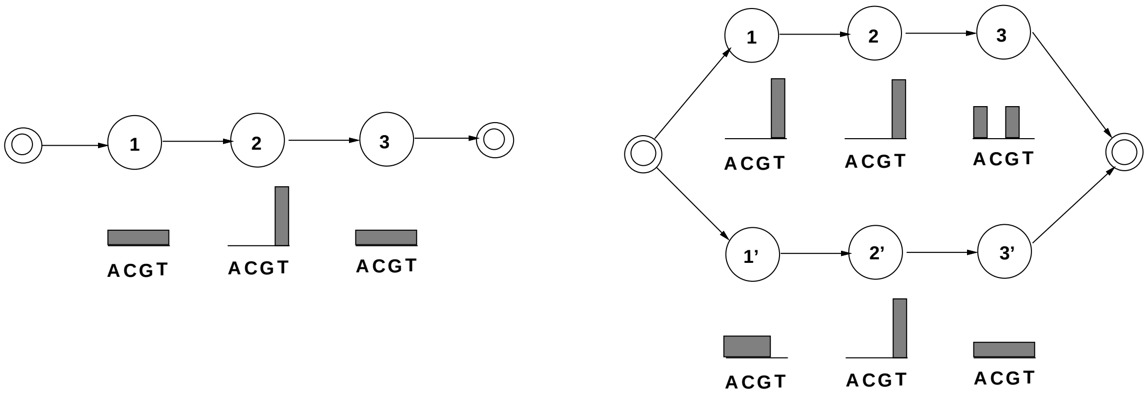
3.1 System overview

The overview of the system is shown in Figure 1. Each component of the diagram is an HMM. Codon HMMs including start codons and stop codons are simple 3 state HMMs. An example of a codon HMM is shown in Figure 3(a). Internal region model is also an HMM, which is a one-state-HMM in current implementation. Each entry of the motif dictionary defines an sequence of amino acids, which are also HMMs. An example of an amino acid HMM is shown in Figure 3(b). While these HMMs are connected as shown in Figure 1, the connected groups in all levels become HMMs because of the modularity described in Section 1.1. For example, the group of start/stop codons, the codon bigram network, the entries of the motif dictionary, the dictionary itself, the group of internal codons, and the whole system are all HMMs.

3.2 Implementation

We used HTK ([15]), which is a commercial software for speech recognition, for the parsing of the DNA sequences. In order to parse the DNA sequences by HTK, we wrote the HMMs, the dictionary and the grammar in the formats of HTK.

The parameters of HMMs, which represent the stochastic features of the various components on the DNA sequences, have been decided from the statistical analysis by the authors. Some of the obvious parameters, for example the probabilities of codon HMMs, have been decided by



```
  ~h "H"
<BeginHMM>
<NumStates> 5
   <State> 2
     ~s "baseX"
   <State> 3
     ~s "baseT"
   <State> 4
     ~s "baseX"
  <TransP> 5
0.0 1.0 0.0 0.0 0.0
0.0 0.0 1.0 0.0 0.0
0.0 0.0 0.0 1.0 0.0
0.0 0.0 0.0 0.0 1.0
0.0 0.0 0.0 0.0 0.0
  <EndHMM>
```

```
           ~h orLIVM
          <BeginHMM>
          <NumStates> 8
             <State> 2
               ~s "baseT"
             <State> 3
               ~s "baseT"
             <State> 4
               ~s "baseR"
             <State> 5
               ~s "baseV"
             <State> 6
               ~s "baseT"
             <State> 7
               ~s "baseX"
            <TransP> 8
0.0 0.5 0.0 0.0 0.5 0.0 0.0 0.0
0.0 0.0 1.0 0.0 0.0 0.0 0.0 0.0
0.0 0.0 0.0 1.0 0.0 0.0 0.0 0.0
0.0 0.0 0.0 0.0 0.0 0.0 0.0 1.0
0.0 0.0 0.0 0.0 0.0 1.0 0.0 0.0
0.0 0.0 0.0 0.0 0.0 0.0 1.0 0.0
0.0 0.0 0.0 0.0 0.0 0.0 0.0 1.0
0.0 0.0 0.0 0.0 0.0 0.0 0.0 0.0
           <EndHMM>
```

(a) HMM of 'orLIVMF'

(b) HMM of 'orLIVM'

Figure 4: Examples of HMMs 'or' patterns of amino acids: the network expression and the definition file.

....	....	....
....	....	....
PS00029	[LEUCINE ZIPPER]	L X X X X X X L X X X X X X L X X X X X X L
PS00039	[DEAD ATP HELICASE]	orLIVMF orLIVMF D E A D orRKEN X orLIVMFYGSTN
PS00054	[RIBOSOMAL S11]	orDNE V T P X orPA X orDN
PS00074	[GLFV DEHYDROGENASE]	orLIV X X G G orSAG K X orGV X X X orDNS orPL
PS00093	[N4 MTASE]	orLIVMF T S P P orFY
PS00097	[CARBAMOYLTRANSFERASE]	F X orEK X S orGT R T
PS00138	[SUBTILASE SER]	G T S X orSA X P X X orSTAVC orAG
PS00154	[ATPASE E1 E2]	D K T G T orLIVM orTI
PS00157	[RUBISCO LARGE]	G X orDN F X K X D E
PS00163	[FUMARATE LYASES]	G S X X M X X K X N
PS00168	[TRP SYNTHASE BETA]	orLIVM X H X G orSTA H K X N
PS00171	[TIM]	orAV Y E P orLIVM W orSA I G T G
....	....	....
....	....	....

Figure 5: An example of entries of motif dictionary

hand. Figure 4 shows the examples of the HMM definitions and their corresponding network expressions.

The dictionary is the collection of all words used by the system. The entries of motifs are the typical example of words. We used the 76 motifs listed in Table 1. As a 'word', the motifs are expressed by the sequence of amino acids in the dictionary. Each amino acid is a simple HMM, whose number of states (3 to 6) depend on the codon patterns of each amino acid (Figure 3(b)). The special wild card X is also implemented by one HMM. This HMM has 9 states, which randomly produce three base symbols except the combinations of stop codons. The regular expressions in PROSITE includes the expression like [LIVM], which means that any of the amino acids in the squared bracket can appear in that position. There are 150 such 'or patterns' in the selected 76 motifs. We constructed the HMMs for those 150 'or patterns.' If the number of amino acids in a 'or pattern' is large, the network of the corresponding HMM can be very complicated. In most cases, however, the networks of the HMMs are relatively simple (Figure 4). Figure 5 illustrates examples of motif entries of the dictionary.

We don't attach probabilities to each words, because probabilities are defined on the bigram of the words. We attach probabilities on the transitions between words. Such probabilities are expressed by stochastic network of the words, which is the grammar of the system. Each word can have multiple entries in the dictionary, the bigram is defined not on the each entry but on each word. In current implementation, the codons are also the words, whose sequence of 'phoneme' consist of only one HMM (codon HMM). This enables us to model di-codon usage by a bigram grammar, which is a first order Markov model between words. The di-codon usage is known as an important statistics ([13]).

4 Results and Discussion

Gene recognition was tested for the data described in Section 2.1, using the motif dictionary with 76-word vocabulary. The 76 motifs were expanded into 199 patterns. There are 156 hits (by means of regular expressions) in the data, all of which are correctly annotated by our recognition system with the dictionary. The system recognized the annotated coding regions in accuracy of around 90% in the scale of base count. Although the annotation of the motifs during the gene recognition was successful, there was no significant improvement of the accuracy of the recognition from the previous work. That was actually unavoidable, because most of the coding regions which includes the motifs in the dictionary had been already recognized correctly by the previous system.

We can construct such stochastic models of proteins in several ways. One simple way is to adopt Markov models on the amino acid sequences. Although a bigram was the limit for the codon usage, trigram or higher Markov models are available for general protein models. That is because the alphabet of amino acids (20) is much smaller than that of codons (64) and because general protein database contains more statistics than the genes of specific organism.

The simplest way is, however, to have the entries of protein database (such as SWISS-PROT) themselves as the models of proteins. If there are exact matches, we can annotate the regions as the matching proteins. Because exact matches are too restrictive for the small changes of amino acids, we can adopt HMMs of similar proteins as the stochastic protein models. Constructing the protein HMMs for the large protein data base is an expensive task in time. We put this for future work. Instead of having the whole protein model, we have adopted the local protein model, that is the motifs, as the stochastic protein model for the gene recognition.

It is necessary to increase the size of vocabulary of the motif dictionary in order to improve the recognition accuracy. Because the construction of the motif dictionary from the motif database is easy, it is not difficult to increase the 'vocabulary' by automatic conversion. However, larger size of the dictionary requires more computational time. Therefore, it is important to select the motifs and their stochastic representation carefully.

It is also necessary to build more precise stochastic models for the motifs. The motif entries in our dictionary are the automatic conversion from the regular expressions of PROSITE entries. Use of the stochastic motifs extracted from the DNA/protein sequences ([4]) is preferable.

It is a future work to build more precise model for the intergenic region model. It should be divided into several models, such as promoters, enhancers and other signals. Using the method of the authors [13] is an obvious next step. We can also build tRNA HMMs from the alignments of tRNA. Because tRNA have context free dependency between their bases, it may be necessary to use stochastic context free grammar (SCFG, [12]).

5 Conclusions

We proposed a new method for using the information of motif database for the recognition of genes in the DNA sequences. The motifs are represented as the 'words' in the motif dictionary, and each 'words' are expressed by the sequence of 'phonemes', which is the HMMs of amino acids on the alphabet of 'A','C','G','T'. The system works just as the speech recognition system, parsing the DNA sequences into 'words'. As a result, this system annotates the position of the motifs, which is defined in the dictionary, in the protein coding regions. The proposed gene recognition system succeeded to annotate the motifs in the genes of DNA sequence of cyanobacterium. In order to improve the recognition accuracy, it is necessary to increase the size of motif dictionary and to make more precise model for the motifs. Implementation of more detailed models for intergenic regions is also necessary.

Acknowledgment

This work was supported in part by Grant-in-Aid for Science Research on Priority Areas, "Genome Science," of the Ministry of Education, Science and Culture of Japan, and in part by Fundamental Research of Science and Technology Agency of Japan. The authors thank Dr.Otsu, the director of Machine Understanding Division and the members of Genome Informatics Group of Electrotechnical Laboratories for the support and the discussions.

References

[1] Asai,K; Hayamizu,S and Onizuka,K.: "HMM with Protein Structure Grammar," *Proceedings of 26th HICSS*, Vol.1, pp. 783-791, 1993.

[2] Borodovsky,M.; Rudd,K.E. and Koonin,E.V.: "Intrinsic and extrinsic approaches for detecting genes in a bacterial genome," *Nucleic Acids Res.*, Vol.22, pp. 4756-4767, 1994.

[3] Dong,S. and Searls,D.B.: "Gene structure prediction by linguistic methods," *Genomics*, Vol.23, pp. 540-551, 1994.

[4] Fujiwara,Y.; Asogawa,M. and Konagaya,A.: "Stochastic Motif Extraction Using Hidden Markov Model," *Proceedings 2nd ISMB*, pp. 121-129, 1994.

[5] GenBank. Genetic sequence data bank, release 92.0. *Technical report, BBN Laboratories, U.S.A.*1995.

[6] Hirosawa,M.: Cyanobase, http://www.kazusa.or.jp/cyano/cyano.html.

[7] Hirosawa,M.; Kaneko,T.; Tabata,S.; McIninch,J.D.; Hayes,W.S.; Borodovsky,M. and Isono,K.: "Computer survey for likely genes in the one megabase contiguous genomic sequence data of *Synechocystis* sp. strain PCC6803," *DNA Res.*, Vol.2, pp. 239-246, 1995.

[8] Kaneko,T.; Tanaka,A.; Sato,S.; Kotani,H.; Suzuki,T.; Miyajima,N.; Sugiura,M. and Tabata,S.: "Sequence analysis of the genome of the unicellular cyanobacterium *Synechocystis* sp. strain PCC6803 sequence features in the 1Mb region from map positions 64% to 92% of the genome," *DNA Res.*, Vol.2, 153-166, 1995.

[9] Krogh,A.; Mian,I.S. and Haussler,D.: "A hidden Markov model that finds gene in E.coli DNA," *Nucleic Acids Res.*, Vol.22, pp. 4768-4778, 1994.

[10] Kulp,D.; Haussler,D.;Reese,M.G. and Eeckman,F.H.: "A Generalized Hidden Markov Model for the Recognition of Human Genes in DNA," *Proceedings, 4th ISMB*, pp.134-142, 1996.

[11] PROSITE release 13.0: *http://expasy.hcuge.ch/sprot/prosite.html.*

[12] Sakakibara,Y.; Brown,M.; Mian,I.S.; Underwood,R. and Haussler,D.: "Stochastic context-free grammars for modeling RNA," *Proceedings of 27th HICSS*, Vol.V, pp. 284-293, 1994.

[13] Yada,T. et al: "Extraction of Hidden Markov Model Representations of Signal Patterns in DNA Sequences," *Pacific Symposium on Biocomputing '96*, pp. 686-696, 1996.

[14] Yada,T. and Hirosawa,M.: "Gene Recognition in Cyanobacterium Genomic Sequence Data Using the Hidden Markov Model," *Proceedings, 4th ISMB*, pp. 252-260, 1996

[15] Young,S.; Jansen,J.; Odell,J.; Ollason,D. and Woodland,P.: *The HTK BOOK*, 1995.