Recognition of Human Genes by Stochastic Parsing

Kiyoshi Asai, Katunobu Itou, Yutaka Ueno
Genome Informatics Group, Electrotechnical Laboratories,
1-1-4 Umezono, Tsukuba, 305 Japan
asai, kito, ueno@etl.go.jp

Tetsushi Yada
Japan Science and Technology Corporation
5-3 Yonbancho, Chiyoda-ku, Tokyo 102 Japan
yada@tokyo.jst-c.go.jp

A gene finding system, **GeneDecoder**, based on a parsing technique using a stochastic grammar and a dictionary of *genetic words* is introduced. The structure of human genes are expressed by a stochastic grammar and a dictionary, whose components are the *genetic words* consisting of *genetic phonemes,* built as hidden Markov models (HMMs). The HMMs represent the nucleotide acid bases, the codons, and the amino acids. The *genetic words* in the dictionary are described by the sequence of these HMMs and represent exons, introns, intergenic regions, tRNA regions and signals in DNA sequences. The statistics between these regions are expressed by the grammar, which is a stochastic network of the *genetic words.* Using the same kind of technique of speech recognition by HMMs with a word dictionary and a grammar, the stochastic network of *genetic words* enables the motif dictionary to be used during the parsing of the DNA sequences. At the same time, stochastic features of donor/acceptor sites, information of the di-codon statistics, and other important features are integrated into stochastic scores during the parsing. As a result, while the system parses DNA sequences and finds the exon/intron structures, 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 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 reasonably finds and annotates the motifs in the exons in the DNA sequence of human.

1 Introduction

The progress of the sequencing projects and the resulting large sequence data demand the computational biologists to develop effective systems to detect genes in the DNA sequences. The exact locations of the genes and the splicing patterns are proved by experiments, but if computational gene finding system can predict the genes correctly, time-consuming experiments may be reduced. There have been proposed a number of systems for finding genes, for example, GENMARK [4], FGENEH[27], GeneID[15], GeneParser[28], Genie[21], GRAIL[30],

GeneHacker[32], HMMgene[19,20], GENSCAN[6].

In this paper, we propose a gene finding system,named **GeneDecoder**, using a parsing technique by a stochastic grammar for combining a protein motif dictionary to the 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 in order to understand the structure [9,28]. However, the components and the rules of the *DNA language* behave as though 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 [6,19,20,21,31,32]. Among them, HMMgene[20] developed independently from **GeneDecoder**, has a similar architecture to our system.

In order to build a stochastic *DNA language* by using HMMs, we model the components of the gene structure by HMMs and connect them by the rules which represent 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, the networks of these objects form a new HMM. This means we can parse the whole DNA sequence by the combined models using a dynamic programming algorithm. However, if we build a precise model of genome structures using many components, we may have to adopt a pruning technique for an integrated model, because the full parsing of a large HMM takes a long time. The situation is same in speech recognition, when we use a word dictionary of a large vocabraty. We use word level pruning technique of speech recognition [18] to solve this problem in gene finding.

Because HMMs have some limitations to express the positional correlations of the bases, some of the models are often made by other methods, like artificial neural nets. Generalized HMMs, which allow such non-HMM models behave as a part of stochastic parsing, have been proposed to combine those non-HMM models with stochastic optimization. [21,6].

1.2 Previous works

The authors developed **GeneHacker**, a gene finding system based on HMMs, to detect the protein coding regions of cyanobacterium (prokaryote), and achieved the recognition accuracy 90.7% for coding regions and 88.1% for intergenic regions[32]. This work is continuing for other prokaryote species.

The recognition performance of that work was reasonable, although the implementation of the models was simple. In that system, the parameters of each HMM 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, where the annotations of the phoneme boundaries are used for the training of the phoneme HMMs. The main statistics between HMMs were codon bigrams, which is a first order Markov model (not *hidden* Markov model) of codons, i.e. using hexamer information by *reading frames* of length three. 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. In order to utilize that property, it is popular to pre-process the sequence by homology search of protein sequences. We have adopted different approach, using protein motifs as the models of these proteins[3], in order to prepare for more flexible models of what the coded proteins are.

1.3 GeneDecoder

While continuing to increase the target of **GeneHacker**, prokaryote genomes, we have expanded the system to model the exon/intron structures of eukaryote genes. **GeneDecoder**, a gene recognition system for eukaryote genes, has been built by adding several components, such as donor/acceptor site models and intron models, and by expanding the stochastic grammar. We have also implemented motif dictionary to **GeneDecoder**, while a protein motif is not necessarily exists within each exon.

2 Data

2.1 DNA sequences

We used the multi-exon part of the non-redundant sets of human genes constructed by Kulp and Reese (1996) as the training/testing sets. For the purpose of comparing the performance with other methods, we also evaluated the system by the test set of 570 vertebrate sequences constructed by Burset and

Guigó[7].

The multi-exon part of the Kulp/Reese data set consists of 9 subsets for the training and the evaluation of gene finding systems. We tested the system for one subset of the data, by the parameters trained from the remaining 7 subsets, excluding the testing subset. The data of training sets determine the HMM parameters and the test sets were used to validate the recognition ability of the **GeneDecoder** whose component HMMs are based on parameters derived from the training set.

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[23], and selected 933 motif patterns as the *genetic 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.

3 System

3.1 System overview

The overview of the system is shown in Figure 1. Each component of the diagram is a *genetic word* or its component *phoneme* HMM. Each component HMM produces the symbols of 'A','C','G','T' during its state transitions. The diagram shows the *grammar* of eukaryotic genes.

The state transition normally begins with intergenic model, then to the start codon, first codon, internal codons (the codon bigram and the motif dictionary). The reading frames are considered by three types of exon fragments at donor/acceptor sites. The intron model behaves differently in three contexts, depending on how many bases of the incomplete codon are placed in the fragment of the exon at the donor site. For example, if the number of fragment bases are two at donor site, the intron model behaves as type two, which is followed only by the fragment of one base at the acceptor site. If a fragment of one base at the donor site, a fragment of two bases at the acceptor site.

While the internal codons are usually come from the codon bigram, the motif dictionary is an alternative path in coding region. By using a motif dictionary as a component of the internal codon model, the motif names are annotated on the predicted coding regions.

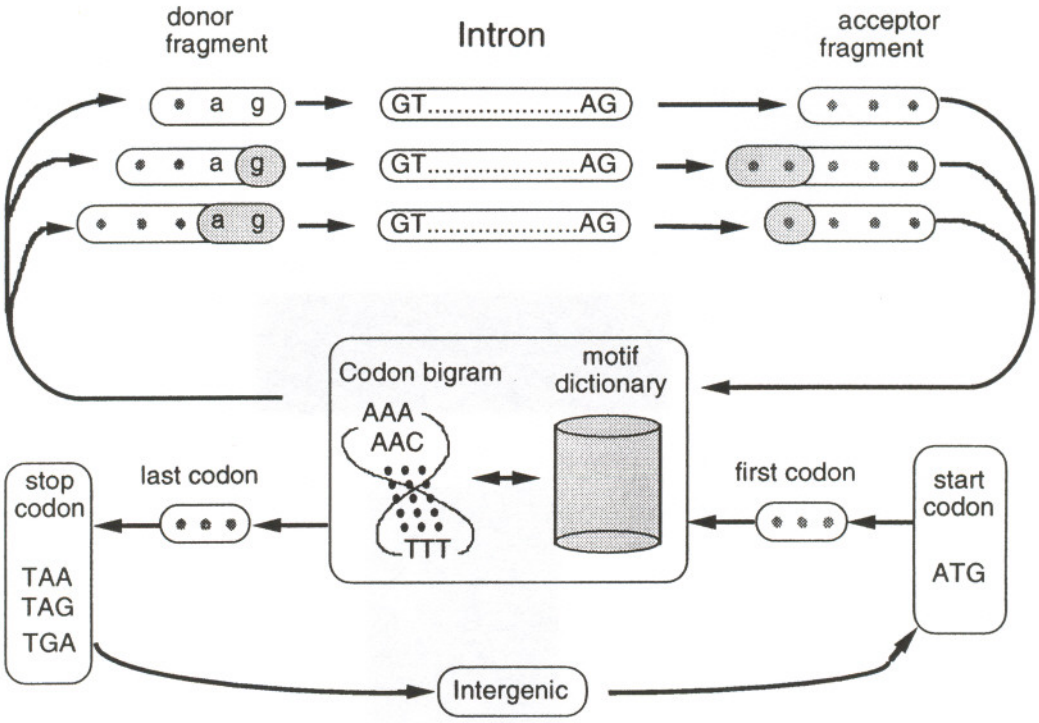


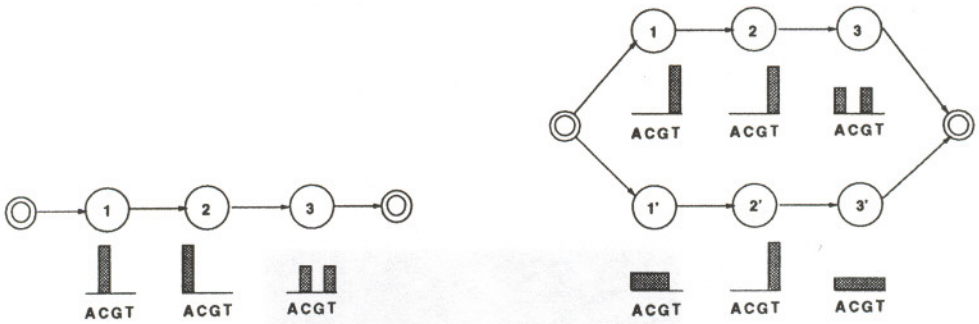
Figure 1: Overview of GeneDecoder

3.2 Exon model: codon bigram and dictionary

The exon model consists of several *genetic words,* internal-codon, donor/acceptor fragments which have three types each, and the motifs. The internal-codon model consists of one HMM, the codon bigram. It consists of 61 three-state blocks, which correspond to internal codons, and these blocks are mutually connected for transition to form a bigram of the codons. The di-codon usage is known as an important statistics[31]. The donor/acceptor fragments have three types, according to the *reading frames* of the exons. They could have the lengths of zero/one/two in base. However, because distributions of the base letters in these positions are very specific, the lengths are set to three/four/five in base.

The motif models consist of motif entries in the dictionary, where each motif entry is a sequence of amino acid models (HMMs), which is similar to the regular patterns in PROSITE. Figure2 shows examples of HMMs of an amino acid and an 'or' pattern of amino acids, which are the *genetic phonemes* in the dictionary. Although each motif has different sequence of *genetic phonemes,* they are all treated as one node in the stochastic grammar. This reduces the cost of the parsing in time.

Figure 1 illustrates examples of entries of the dictionary.



(a) HMM of amino acid Histidine (b) HMM of 'orLIVM'

Figure 2: Examples of HMMs of an amino acid and 'or' patterns of amino acids: The states expressed by the double circles are special state of null output.

Table 1: An example of a dictionary

genetic word	sequence of genetic phonemes
5'intergenic	5'intergenic-main 5'intergenic-tail
3'intergenic	3'intergenic-head 3'intergenic-main
start-part	startcodon firstcodon
internal-codon	codon-bigram
stop-part	lastcodon stopcodon
exon-donor-fragment3	exon-donor-fragment3
exon-donor-fragment4	exon-donor-fragment4
exon-donor-fragment5	exon-donor-fragment5
exon-acceptor-fragment3	exon-acceptor-fragment3
exon-acceptor-fragment5	exon-acceptor-fragment5
exon-acceptor-fragment4	exon-acceptor-fragment4
intron	intronhead introncenter introntail
intron	intronhead introncenter branch-point branch-tail
LEUCINE ZIPPER	L X X X X X X L X X X X X X L X X X X X X L
DEAD ATP HELICASE	orLIVMF orLIVMF D E A D orRKEN X orLIVMFYGSTN
RIBOSOMAL S11	orDNE V T P X orPA X orDN
GLFV DEHYDROGENASE	orLIV X X G G orSAG K X orGV X X X orDNS orPL
N4 MTASE	orLIVMF T S P P orFY
CARBAMOYLTRANSFERASE	F X orEK X S orGT R T
SUBTILASE SER	G T S X orSA X P X X orSTAVC orAG
ATPASE E1 E2	D K T G T orLIVM orTI
RUBISCO LARGE	G X orDN F X K X D E
FUMARATE LYASES	G S X X M X X K X N
TRP SYNTHASE BETA	orLIVM X H X G orSTA H K X N
TIM	orAV Y E P orLIVM W orSA I G T G
.....	.....
.....	.....
.....	.....
.....	.....

3.3 Intron model and splice site

While the intron model behaves three ways, the intron model itself has fixed sequence of *genetic phonemes,* as is listed in Table1. It consists of intronhead, introncenter and introntail. Intronhead and the introntail are the fixed-length HMMs, which are equivalent of weight matrices of the probabilities, whose lengths are 7 and 17 respectively. Introncenter is a complex of a third order Markov model and a fixed-length HMMs which include branch point model. The third order Markov model is implemented as an HMM, by assigning different states for all second order contexts of the bases. The pattern of the branch point model was built by the motif extraction module of YEBIS [31]. It works as follows: (1) selection of significant subsequences; (2) classification of the subsequences into groups of patterns; (3) assignment of characteristic base length for each group; (4) determination of HMM for each group; Finally, the method can generate HMMs each of which describe a pattern in the data. The resulting branch point model is left-to-right HMMs of six states, whose consensus pattern is 'CCTGAC'.

GeneDecoder has no separate splice site model. Splice site statistics are distributed in the donor/acceptor fragment models of exons and the fixed length part of the intron models. Because they are left-to-right HMMs, it can be said that **GeneDecoder** is using a weight matrix for the detection of the candidate of the splice site.

3.4 Intergenic model

The intergenic model is a combination of fixed length model adjacent to first/last exons and a third order Markov model. The implementation of these models are almost same as the intron model.

3.5 Training

The training of the models has been performed simply calculating the statistics of the annotated regions of the training data set. No time-consuming Baum-Welch training of the parameters of HMMs is necessary.

3.6 Parsing

The recognition process is exactly the same as the dynamic programming parsing of the speech, using a grammar defined on these *genetic words.* We use word level pruning and N-best parsing techniques[18] for this parsing. Having 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, but the former uses the information of the database *during* the process of deciding the candidate of the coding regions.

4 Results

The recognitions results for Burset/Guigó set of 570 vertebrate genes are shown in Table2. The training set was chosen from Kulp/Reese set of human genes. The data is not completely independent, so human genes tend to be predicted with high accuracies. The performance of **GeneDecoder** is not higher than GENSCAN, HMMgene and newer version of Genie. Except these new gene finding systems, GeneDecoder performs relatively well, although the implementation is quite simple and not using homology searches by protein sequences. Among those 570 genes, 123 genes were correctly predicted with their complete exon/intron structures.

Table3 shows the performance for Kulp/Reese set of human genes.

5 Discussion

Gene recognition by **GeneDecoder** was tested for the data described in Section 2.1, using the motif dictionary with 933-word vocabulary. In the coding sequences of Burset/Guig'o data, which are annotated as 'CDS' in GenBank, there are 241 hits of 97 motifs using the *dictionary* described in 2.2. Among them, 167 hits are correctly annotated by **GeneDecoder**. There are 74 hits are missed by the system because those motifs are not completely included in the exons. Because **GeneDecoder** searches the motifs only within the exons, it could not be avoided using current set of motif dictionary. We may need to construct a new motif dictionary, which has the entries of the protein motifs within the exons. There are also 79 false hits of motifs because of the miss-alignments of the exon/intron structures. Because a hit of a motif forms strong bias for the region to be a part of the exon, those false hits prevent the improvement of the recognition accuracy by the motif dictionary. Further study is needed to solve this problem.

While the motif dictionary itself should be improved, an obvious extension of the motif dictionary is having cDNA data as the entries of the dictionary. Having the protein database as the entries of the dictionary, however, may not be a good approach for using homology information of proteins. Homology searches can be performed as pre-/post- process of the system. By pre-processing, we get scores of homology search for each region and we can

Table 2: Performance comparison for Burset/Guigó set of 570 vertebrate genes

	Accuracy per nucleotide				Accuracy per exon				
Program	Sn	Sp	AC	CC	Sn	Sp	Avg.	ME	WE
GeneDecoder	0.87	0.82	0.81	0.81	0.62	0.51	0.57	0.13	0.11
GENSCAN	0.93	0.93	0.91	0.92	0.78	0.81	0.80	0.09	0.05
HMMgene	0.88	0.94	n/a	n/a	0.74	0.78	n/a	0.13	0.08
Genie	0.87	0.88	0.85	n/a	0.69	0.70	0.69	0.10	0.15
GeneID	0.63	0.81	0.67	0.65	0.44	0.46	0.45	0.28	0.24
GenLang	0.72	0.79	0.69	0.71	0.51	0.52	0.52	0.21	0.22
GeneParser2	0.66	0.79	0.67	0.65	0.35	0.40	0.37	0.34	0.17
GRAIL2	0.72	0.87	0.75	0.76	0.36	0.43	0.40	0.25	0.11
SORFIND	0.71	0.85	0.73	0.72	0.42	0.47	0.45	0.24	0.14
Xpound	0.61	0.87	0.68	0.69	0.15	0.18	0.17	0.33	0.13
GeneID+	0.91	0.91	0.88	0.88	0.73	0.70	0.71	0.07	0.13
GeneParser3	0.86	0.91	0.86	0.85	0.56	0.58	0.57	0.14	0.09

Sn and Sp are the sensitivity and specificity respectively. Accuracy per nucleotide shows base level performance, number of bases which was correctly predicted as exon or non-exon is counted. Accuracy per exon shows exon level performance, where Sn, Sp and Avg are counted for exactly corrected exons only. AC and CC are Approximate Correlation and Correlation Coefficient respectively. For precise definition, see Burge (1997)[6] for example. The result of GENSCAN[6], HMMgene[20] and Genie[22] are from the references. The other results are from Burset and Guigó[7].

Table 3: Performance comparison for Kulp/Reese set of human genes: Tested on only multi-exon part of the data set

	Accuracy per nucleotide				Accuracy per exon				
Program	Sn	Sp	AC	CC	Sn	Sp	Avg.	ME	WE
Part0	0.88	0.76	0.79	0.79	0.60	0.50	0.55	0.13	0.11
Part1	0.90	0.76	0.81	0.81	0.68	0.55	0.61	0.14	0.11
Part2	0.87	0.82	0.82	0.83	0.59	0.48	0.54	0.17	0.14
Part3	0.85	0.74	0.76	0.76	0.51	0.42	0.56	0.21	0.17
Part4	0.92	0.87	0.87	0.87	0.64	0.58	0.61	0.12	0.11
Part5	0.82	0.66	0.70	0.77	0.50	0.37	0.49	0.19	0.14
Part6	0.69	0.57	0.61	0.60	0.44	0.33	0.39	0.29	0.22
Part7	0.94	0.72	0.80	0.79	0.65	0.50	0.57	0.06	0.04
Part8	0.90	0.63	0.74	0.73	0.61	0.42	0.52	0.11	0.07
Average	0.87	0.72	0.77	0.77	0.58	0.45	0.52	0.16	0.12
Genie	0.74	0.81	0.74	n/a	0.59	0.59	0.59	0.17	0.21
HMMgene	0.82	0.94	n/a	n/a	0.64	0.79	n/a	0.23	0.06

integrate these scores into the stochastic parsing. On the other hand, parsing with N-best prediction enables homology searches to be post-processes.

The splice site sensor is equivalent to a weight matrix in current implementation. It may be necessary to adopt more complicated sensors, such as artificial neural nets. It can be also integrated into **GeneDecoder** by a pre-/post- process technique.

The promoter model is under construction. The authors have proposed a method for such DNA motif construction[31].

6 Conclusions

We have developed a gene recognition system, named **GeneDecoder**, which consists of HMM-based components and a stochastic grammar. The performance of the recognition is reasonable as a system without homology search.

The motif patterns of PROSITE was used as the entries of the dictionary. The motifs are represented as the *genetic words* in the motif dictionary, and each *genetic 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 *genetic words.* As a result, this system annotates the position of the motifs, which is defined in the dictionary, in the protein coding regions.

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 Real World Computing (RWC) Program of the Ministry of International Trade and Industry 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.; Handa,K. and Hayamizu,S.: "Genetic Information Processing by Stochastic Model: HMM for Secondary Structure Prediction of Protein," *Genome Informatics,* **2**, 144-147 (in Japanese, 1991).
2. Asai,K.; Hayamizu,S. and Onizuka,K.: "HMM with Protein Structure Grammar," *Proceedings of 26th HICSS*, **1**, 783-791 (1993).

3. Asai,K.; Yada,T. and Itou,K.: "Finding Genes by Hidden Markov Models with a Protein Motif Dictionary," *Genome Informatics,* **7**, 88-97 (1996).
4. Borodovsky,M. and McIninch,J.: "GENMARK: parallel gene recognition for both DNA strands," *Comp. Chem.* **17**, 123-133 (1993).
5. Bucher,P.: "Weight matrix descriptions of four eukaryotic RNA polymerase II Promoter elements derived from 502 unrelated promoter sequences," *J.Mol.Biol.* **202**, 563-578 (1990).
6. Burge,C. and Karlin S.: "Prediction of Complete Gene Structures in Human Genomic DNA," *J.Mol.Biol.* **268**, 78-94 (1997).
7. Burset,M and Guigó,R.: "Evaluation of gene structure prediction programs," *Genomics*, **34**, 353-367 (1996).
8. Fickett,J.: "Assessment of protein coding measures," *Nucl. Acids Res.* **20**, 6441-6450 (1992).
9. Dong,S. and Searls,D.B.: "Gene structure prediction by linguistic methods," *Genomics*, **23**, 540-551 (1994).
10. Fujiwara,Y.; Asogawa,M. and Konagaya,A.: "Stochastic Motif Extraction Using Hidden Markov Model," *ISMB94*, **2**, 121-129 (1994).
11. Gelfand,M.S. and Roytberg,M.A.: "Prediction of the intron-exon structure by a dynamic programming approach," BioSystems, **30**, 173-182 (1993).
12. Gelfand,M.S.: "Prediction of function in DNA sequence analysis," *J. Comp. Biol.* **2**(1), 87-115 (1995).
13. Gelfand,M.S.; Mironov,A.A. and Pevzner,P.: "Gene recognition via spliced alignment," *Proc. Natl. Acad. Sci. USA*, **93**, 3015-3019 (1996).
14. GenBank. Genetic sequence data bank, release 92.0. *Technical report, BBN Laboratories, U.S.A.* (1995).
15. Guigó,R.; Knudsen,S.; Drake,N. and Smith,T.: "Prediction of gene structure," *J. Mol. Biol.* **226**, 141-157 (1992).
16. Henderson,J.; Salzberg,S. and Fasman,K.: "Finding Genes in DNA with a Hidden Markov Model," *J. Comp. Biol.* **4**(2), 127-141 (1997).
17. Hutchinson,G.B. and Hayden, M.R.: "The prediction of exons through an analysis of spliceable open reading frames." *Nucleic Acids Research, 20:13* 3453-3462(1992).
18. Itou,K.; Hayamizu,S. and Tanaka,H.: "Continuous Speech Recognition by Context-Dependent Phonetic HMM and an Efficient Algorithm for Finding N-best Sentence Hypotheses," *ICASSP-92*, I-21–24 (1992).
19. Krogh,A.; Mian,I.S. and Haussler,D.: "A hidden Markov model that finds gene in E.coli DNA," *Nucleic Acids Res.*, **22**, 4768-4778 (1994).
20. Krogh,A.: "Two methods for improving performance of an HMM and their application for gene finding," *ISMB97*, **5**, 179-186 (1997).