Detection of Exons with Deletions and Insertions by Hidden Markov Models*

YANG Wen-Qiang, QIAN Min-Ping**
(*The School of Mathematics, Peking University, Beijing* 100871, China)
HUANG Da-Wei
(*Bell Labs Research China, Lucent Technologies, Beijing* 100080, China)

Abstract After more and more genome sequencing projects, like the “Human Genome Project", the prediction of genes, including their coding region and their regulatory region, has received a lot of attention. Softwares such as GENSCAN and GeneMark are powerful, but still do not meet the requirement of the practical application. The GENSCAN predicts exons accurately, if the sequences predicted does not have insertions and deletions in their coding regions. But if it does have, even only one, the prediction could be disturbed seriously and satisfactory results can not be obtained. A hidden Markov model with states of deletions, insertions and main state is introduced to find the error of deletions and insertions. The result shows that sensitivity and specificity in exon level are both higher than 84% on the *Burset/Guigò* test data set.

Key words gene finding, hidden Markov model, Viterbi algorithm

1 Introduction

The “Human Genome Project" is close to be completed, and many other genome sequencing projects, such as the mouse and the rice are being carried out. Finding gene from uncharacterized genomic sequence by using computational tools unquestionably has practical interest. To locate the gene accurately by computational methods will simplify the analysis of large uncharacterized sequence data and will speed up the pace of projects after genome sequencing.

The problem is relatively simple for prokaryotic DNA sequences, since their protein coding region has no introns and their open reading frames(ORF) are continuous. It is more difficult for eukaryotic DNA sequences due to the presence of introns between the relatively short exons. Thus to discriminate exons and introns reliably may be more difficult and complicate. Especially, when there are insertions and deletions in the exons, the open reading frame will be disturbed, and the judgement of whether the segments being detected have the property of CDS sequences becomes confusing.

Methods of predicting potential protein coding regions in genomic sequences have developed since 1980s. Many efforts have been devoted to this aim and big progress has been achieved. There are a number of computer programs for gene identification, including SORFIND[1], GeneID[2], GeneMark[3], Xpound[4], FGENEH[5], GRAIL2[6], GeneParser[7], Genie[8], GeneWise[9], GENESCAN[10], INFO[11], Procrustes[12, 13]. Most of these programs make use of sophisticated pattern recognition techniques such as linear discriminant analysis, neural networks(NN), or Hidden Markov Models(HMM) to identify coding regions. For example: GENSCAN (see [10]), the most successful method in recent years, uses a general HMM, where the method of Maximal Dependence Decomposition is introduced to model the donor splice signal. The predicting accuracy of GENSCAN is better than most other gene finding programs. GeneMark is also based on HMM, and was tested on the *E. coli* complete genome with most genes being identified. Burset and Guigó compare many of these programs with a test set of 570 whole gene sequences by using several accuracy measures. The average specificity and sensitivity at exon level varies from 0.17 to 0.63 (see [14] and [15]).

The accuracy of those programs is good in DNA sequences free of errors. But sequences newly submitted to the programs, however, will often contain artificial nucleotide insertions and deletions. It is pointed out in [14], that the accuracy of those programs will be low when there are insertions or deletions in the coding regions of DNA sequences.

In this paper, we try to find the exons with a few deletions or insertions on an algorithm based on HMM. It is well known that the statistical features of coding regions in the right open reading frame are different from that in noncoding regions, while for the wrong open reading frames they are not. There

* This work was supported by grants from the National Natural Sciences Foundation of China (19971005), DPF1 of Higher Education Program and Bell Lab Research China.

** Corresponding author.

Tel: 86-10-62752525, E-mail: qianmp@math.pku.edu.cn

Received: April 19, 2001 Accepted: July 9, 2001

are fairly strong preferences of codons and transition from codon to codon. Such preferences could be disturbed seriously when there are insertions or deletions in coding regions and the open reading frame is broken. To take advantages of this statistical property of coding regions, we define scores for each sliding window in a given sequence, with the consideration of possible insertions and deletions. In fact, when a deletion (insertion) appears in a sliding window, the score becomes low, while the score under the open reading frame one base-pair shifted to the left (right) will keep as high as before. Thus we use a score system based on this phenomenon to get the observation processes in the HMM.

2 Hidden Markov Model (HMM)

The HMM has been successfully applied to pattern recognition problems , such as the speech recognition (see [16]) and gene finding (see [10]). In this paper the mathematical tool of HMM is applied to uncover the pattern of transition from codon to codon and to find the error of deletions and insertions in exons.

A HMM is a pair of stochastic processes: an underlying Markov chain (or field) and an observable stochastic process (or field). The Markov chain is not observable, and can only be understood through the observable process.

The Markov chain (or field) in HMM is essentially a collection of finite states connected by transitions with the Markov property. The structure of the hidden Markov model that we used here isillustrated in figure 1. The states of the Markov chain $\{X_n\}_{n\geqslant 1}$ is $\{M_1, M_2, M_3, I, D\}$, where M_1, M_2, M_3correspond to the first, second, third codon position respectively. The state D represents the deletion and I represents the insertion. For convenience, we have added a start state S and an end state E. Namely the state space of the HMM is $\{M_1, M_2, M_3, I, D, S, E\}$.

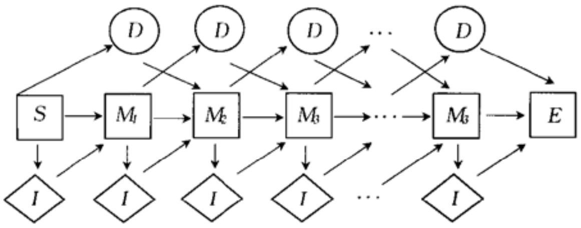

Fig. 1 The hidden Markov model

Let us consider the probability space $(\Omega, F, \{F_n\}, P)$ and denote the Markov chain with the states space $S = \{s_1, s_2, \ldots, s_m\}.$, by $(X_n)_{n\geqslant 1}$, while the observable stochastic process by $(Y_n)_{n\geqslant 1}$, on the space $(\Omega, F, \{F_n\}, P)$.

Assume that

$P(Y_n = y_n \mid X_n = x_n, X_{n-1}, \ldots, X_1, Y_1, \ldots, Y_{n-1}) = P(Y_n = y_n \mid X_n = x_n)$

Let $A = \{a_{ij}\}$ be the transition probability matrix of the Markov chain, i. e.

$a_{ij} = Pr\{X_{n+1} = j \mid X_n = i\}$

The emission probability matrix $B = \{b_j(O_n)\}$ is defined as

$b_j(O_k) = Pr\{Y_n = O_k \mid X_n = j\}$

We denote the initial distribution of the process $(X_n)_{n\geqslant 1}$, by $\Pi = \{\pi_i\}$, namely

$\pi_i = Pr\{X_0 = s_i\}$

Then the hidden Markov model can be expressed as

$\lambda = (\Pi, A, B)$

Noticing that the probability of the appearance of a nucleotide is different in each of the three positions of a codon. Let $f^1(b, i)$ be the probability of the nucleotide b at the codon position i; $(i = 1, 2, 3)$; $f_j^2(b_2, b_1)$ be the conditional probability of nucleotide b_2 at the codon position $j+1$, given the nucleotide b_1 at the codon position j $(j = 1, 2)$; and $f^3(b_3, b_2, b_1)$ be the conditional probability of the nucleotide b_3 at the codon position 3, given the nucleotide b_2 at the codon position 2 and b_1 at the codon position 1. For a given DNA sequence $C = c_1c_2\ldots c_n$, there are three possible open reading frames:

Frame 1 starting from c_1:

$c_1c_2c_3, c_4c_5c_6, \ldots$

Frame 2 starting from c_2:

$c_1, c_2c_3c_4, c_5c_6c_7, \ldots$

Frame 3 starting from c_3:

$c_1c_2, c_3c_4c_5, c_6c_7c_8, \ldots$

The probability of C, under the j-th read frame $(j = 1, 2, 3)$, can be calculate respectively as

$$Pr^1(C) = f^1(c_1, 1)f_1^2(c_2, c_1)f^3(c_1, c_2, c_1)f^1(c_4, 1)f_1^2(c_5, c_4)f^3(c_6, c_5, c_4)\ \ldots$$
$$Pr^2(C) = f^1(c_1, 3)f^1(c_2, 1)f_1^2(c_3, c_2)f^3(c_4, c_3, c_2)f^1(c_5, 1)f_1^2(c_6, c_5)f^3(c_7, c_6, c_5)\ \ldots$$
$$Pr^1(C) = f^1(c_1, 2)f_2^2(c_2, c_1)f^1(c_3, 1)f_1^2(c_4, c_3)f^3(c_5, c_4, c_3)f^1(c_6, 1)f_1^2(c_7, c_6)f^3(c_8, c_7, c_6)\ \ldots$$

The *log-score* of the sequence C under the j-th open reading frame $(j = 1, 2, 3)$ is defined as

$$\textit{log-score}^j(C) = \max_j\{\log_2 Pr^j(C) - 2n\}$$

To illustrate the *log-score*, we use a 1 200 bp

DNA sequence from the *gamma-globin* gene of the human genome, extracted from EMBL, entry AGGGLINE from 3 000 to 4 200. This gene has three exons located at 3 066~ 3 157, 3 281~ 3 503 and 4 393 ~ 4 521 (not shown infigure 2) respectively. The plots *log-score*^{3} of this sequence without and with one deletion at 3360 are given in the figure 2 and figure 3 respectively, in sliding window of length 60 bp. One can see that in figure 1, the *log-score*^{3} reflects the exons fairly well, while in figure 3, after the deletion the *log-score*^{3} becomes low.

In the present paper, the Baum-welch algorithm and the EM algorithm are used to estimate iteratively training for getting parameters. To detect the rough positions of insertions and deletions the Viterbi algorithm is used. The accurate splicing sites of donors and accepters are determined the by the technique similar to GENSCAN (see [10]).

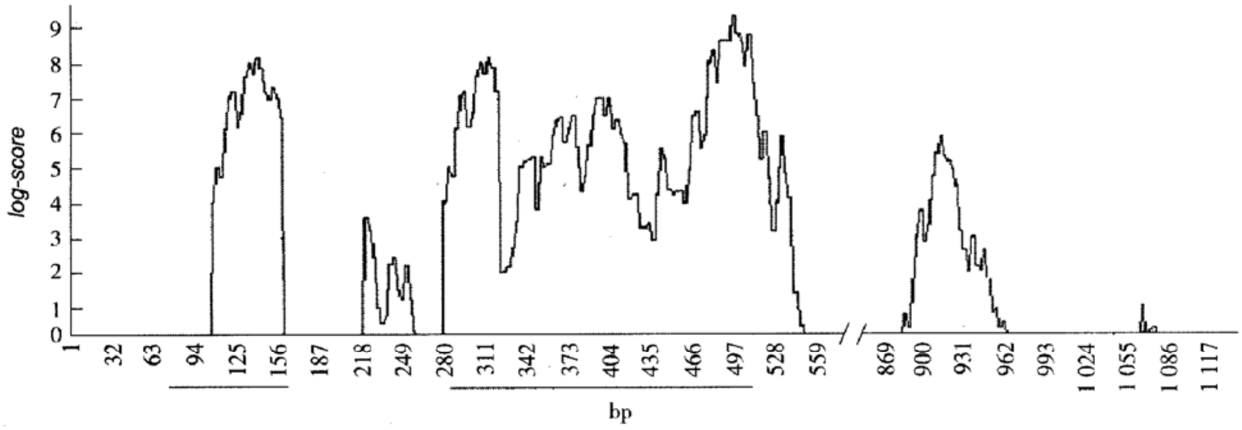

Fig. 2 The *log-score* of DNA sequence segment of AGGGLINE

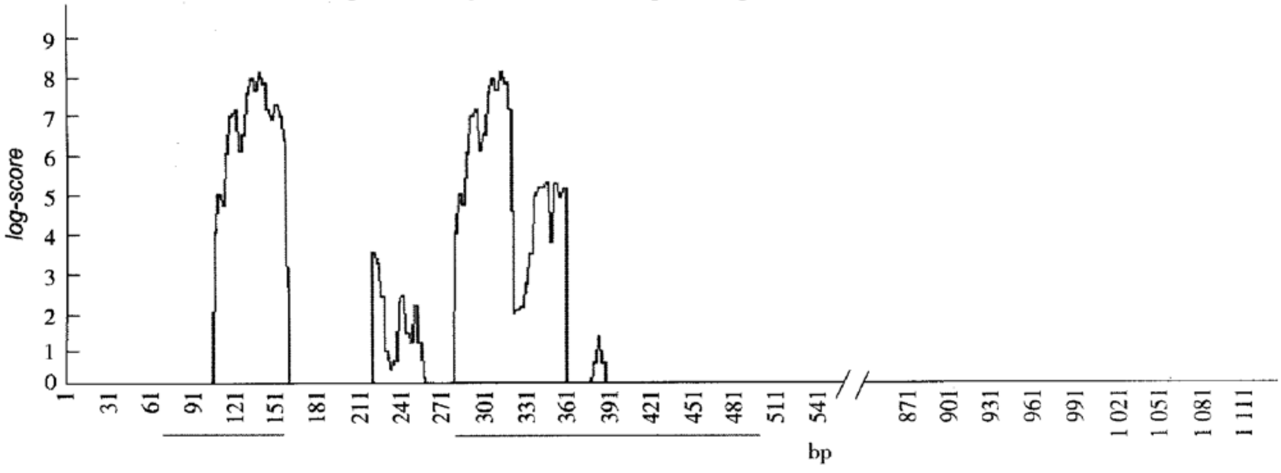

Fig. 3 The *log-score* of DNA sequence segment of AGGGLINE with one deletion in the exon

3 Data and Result

The data for learning in our experiment are prepared as follows. We extract a set of 400 whole human genes from the GenBank released 112. Then we add deletions and insertions in the coding exons randomly both at the rate of 0.3%. This is based on the research of Koop *et al* (see [17]), which concludes that error rates for four kinds of errors (mismatch, ambiguities, insertions and deletions) remains fairly constant and under 1%. We choose the set of 570 vertebrate genes constructed by Burset/ Guigó sequence set, which is considered as a standard clean comparison data for gene predictions, as our test data set. This data set can be accessed through the World Wide Web at the URL:

" http://www. imim. es/GeneIdentification/ Evaluation/Index. html"

Then we delete or insert some bases in some exon region randomly at the rate of 0.3% to produce the test data set with deletions and insertions in the exons.

Four indices commonly used for the accuracy at the exon level are: sensitivity *ESn*, the proportion of exons which are predicted correctly, in actual exons without deletions and insertions, *ME*, the proportion of actual exons without deletions or insertions are predicted to be error exons; the specificity *ESp* and

WE, the proportion of exons predicted correctly and incorrectly respectively in all exons predicted without deletions and insertions (see [4] for review). The result of prediction is shown in the Table 1.

Table 1 The result of the prediction

	Pridicted normal exon	Pridicted error exon
Normal exon	493	77
Error exon	82	437

Thus the accuracy indices *ESn*, *ESp*, *ME*, *WE* are as follows:

$$ESn = 493/(493+77) = 86\%$$

$$ESp = 493/(493+82) = 85\%$$

$$ME = 77/(493+77) = 14\%$$

$$WE = 82/(493+82) = 14\%$$

References

1 Huntchinson G B, Hayden M R. The Prediction of exons through an analysis of spliceable open reading frames. Nucleic Acids Res, 1992, **20**: 3453~ 3462

2 Guigo R, Knudsen N, Smith T F. Prediction of gene strudture. J Mol Biol, 1992, **226**: 141~ 157

3 Borodovsky M, McIninch J, GeneMark: Parallel gene recognition for both DNA strands. Computer & Chemistry, 1993, **17** (2): 123~ 133

4 Thomas A, Skolnick M H. A probabilistic model for detecting coding region in DNA sequence. IMA J Math Appl Med Biol, 1994, **11**: 149~ 160

5 Solovsky V V, Salamov A A, Lawrence C B. Predicting internal exons by oligonucleotide composition and discrinant analysis of spliceable open reading frames. Nucleic Acids Res, 1994, **22**: 5156~ 5163

6 Xu J R, Mural R J, Shah M, *et al*. An improved system for exon recognition and gene modeling in human DNA sequences. Proceedings of the Second International Conference on Intelligent Systems for Molecular Biology, Menlo Park, 1994.

7 Snyder E E, Stormo G D. Identification of coding regions in genomic DNA. J Mol Biol, 1995, **21**: 1~ 18

8 Kulp D, Haussier D, Reese M G, *et al*. A generalized Hidden Markov Model for the recognition of human genes in DNA. Proceedings of the Fourth International Conference on Intelligent Systems for Molecular Biology, Menlo Park, 1996.

9 Birney E, PairWise, SearchWise. Finding the optimal alignment in a simultaneous comparison of a protein profile against all DNA translation frames. Nucleic Acids Res, 1996, **24**: 2730~ 2739

10 Burge C, Karlin S. Prediction of complete gene structures in human genomics DNA. J Mol Biol, 1997, **268**: 78~ 94

11 Laub M T, Smith D W. Finding intron/ exon splice junctions using INFO, interruption finder and organizer. J Comput Biol, 1998, **5**: 307~ 321

12 Gelfand M S, Mironov A A, Pevzner P A. Gene recognition via spliced sequencealignment. Proc Nalt Acad Sci USA, 1996, **93**: 9061~ 9066

13 Mironov A A, Roytberg M A, Pevzner P A, *et al*. Performance-guarantee gene predictions via spliced alignment. Genomics, 1998, **51**: 332~ 339

14 Burset M, Guigo R. Evaluation of gene structure prediction programs. Genomics, 1996, **34**: 353~ 367

15 Guigo R, Agarwal P, Abril J F, *et al*. An assessment of gene prediction accuracy in large DNA sequences. Genome Research, 2000, **10** (10): 1631~ 1642

16 Rabiner L R. Atutorial on hidden Markov models and selected applications in speech recognition. Proc IEEE, 1989, **77**: 257~ 286

17 Koop B F, Chen W Q, *et al*. Sequence length and error analysis of sequence and automated Taq cycle sequencing method. Bio Techniques, 1993, **14** (3): 442~ 447

基于隐马氏模型对编码序列缺失与插入的检测*

杨文强　钱敏平**

(北京大学数学学院, 北京 100871)

HUANG Da-Wei

(贝尔实验室中国基础科学研究院, 北京 100080)

摘要　在基因组测序工作完成后, 利用计算工具进行基因识别以及基因结构预测受到了越来越多人的重视. 人们开发了大量的相关应用软件, 如 GenScan, Genemark, GRAIL 等, 这些软件在寻找新基因方面提供了很重要的线索. 但基因的识别和预测问题仍未得到完全解决, 当目标基因的编码序列有缺失和插入时, 其预测结果和基因的实际结构相差很大. 为了消除测序错误对预测结果的影响, 希望能找出编码序列区的测序错误. 基于这种想法, 尝试根据 DNA 序列的一些统计特性, 利用隐马尔科夫模型 (Hidden Markov Model), 引入缺失和插入状态, 然后用 Viterbi 算法, 从中找出含有缺失和插入的外显子序列片段. 在常用的 Burset/ Guigo 检测集进行检测, 得到的结果在外显子水平上, *Sn* (sensitivity) 和 *Sp* (specificity) 均达到 84% 以上.

关键词　基因识别, 隐马尔科夫模型, Viterbi 算法

学科分类号　Q612

[*] 国家自然科学基金资助 (19971005), 高等学校博士学科点专项科研基金及贝尔实验室中国基础科学研究院资助.

[**] 通讯联系人.　Tel: 010-62752525, E-mail: qianmp@ math. pku. edu. cn　收稿日期: 2001-04-19, 接受日期: 2001-07-09