The Prediction of Human Genes in DNA Based on a Generalized Hidden Markov Model

Rui Guo(✉), Ke Yan, Wei He, and Jian Zhang

Bio-Computing Research Center, Shenzhen Graduate School, Harbin Institute of Technology,
Shenzhen, China
570930945@qq.com

Abstract. The Generalized Hidden Markov Model (GHMM) has been proved to be an excellently general probabilistic model of the gene structure of human genomic sequences. It can simultaneously incorporate different signal descriptions like splicing sites and content descriptions, for instance, compositional features of exons and introns. Enjoying its flexibility and convincing probabilistic underpinnings, we integrate some other modification of submodels and then implement a prediction program of Human Genes in DNA. The program has the capacity to predict multiple genes in a sequence, to deal with partial as well as complete genes, and to predict consistent sets of genes occurring on either or both DNA strands. More importantly, it also can perform well for longer sequences with an unknown number of genes in them. In the experiments, the results show that the proposed method has better performance in prediction accuracy than some existing methods, and over 70 % of exons can be identified exactly.

Keywords: Gene prediction · WWAM · IMM · GHMM · The prefix sum arrays · The method based on similarity weighting of sequence patterns

1 Introduction

In recent years, with the development and gradual promotion of the third-generation gene sequencing technology [1], its sequencing cost becomes lower while each sequencing length becomes much longer and sequencing accuracy much higher, we have been accumulating the higher quality of genome sequences for all kinds of organisms at a faster rate. In order to tap the potential value of these data, a good many gene-finding programs, which identify gene in genomic DNA sequences by computational methods, are routinely used by gene annotation project members to help identify genes in that newly sequenced regions [2].

A complete gene structure in eukaryotes generally consists of some different functional elements which are divided into signal sensors and content sensors [3]. Signal sensors are regarded as the basic method of finding the presence of functional sites with fixed length, such as promoters, start and stop codons, splice sites, branch points, etc. As for content sensors, they are measures that try to classify a DNA region into coding and noncoding. The measures are mainly based on extrinsic similarity with a biologically characterized sequence, e.g., protein sequence, cDNA or expressed sequence tag (EST)

Z. You et al. (Eds.): CCBR 2016, LNCS 9967, pp. 747–755, 2016.
DOI: 10.1007/978-3-319-46654-5_82

sequence, and intrinsic statistical properties such as codon usage(a triplet of DNA bases), hexamer frequency, nucleotide composition, GC content and base occurrence periodicity.

Many early approaches only focused on those signal sensors to roughly locate the position of gene in genomic DNA sequences. Subsequently, in order to predict entire gene structures precisely, the approaches have been developed by integrating multiple types of information which include splice signal sensors, compositional properties of coding and non-coding DNA, and database homology searching in some cases. Some typical programs show: GENEID [4], Genie [5], GENSCAN [2] and AUGUSTUS [6]. However, early available programs have two important limitations [7]: one is that their algorithms assume that the input sequence contains exactly one complete gene. If the sequence contains a partial or multiple genes, the results they provide do not make sense. The other is that due to evaluating by independent control sets, the accuracy is usually worse than originally thought. Fortunately, some methods emerging afterwards supply these gaps, such as GENSCAN and AUGUSTUS. They use an explicitly double-stranded genomic sequence model to simultaneously analyze potential genes occurring on both DNA strands. Additionally, the model treats the general case in which the sequence can contain a partial gene, a complete gene, multiple complete genes, or no gene at all. The combination of double-stranded nature of model and the capacity to deal with variable numbers of genes may prove useful for long human genomic segments, e.g. those of a hundred kilobases or more, which usually contain more than a gene on one or both strands. We follow the model design, integrate some other different innovations of submodels and implement a prediction system of Human Genes in DNA. The system has functional advantages mentioned above and a high performance in accuracy.

Finally, regardless of benefits of function and performance in our model, the difficulties of handling overlapping transcription units and explicitly addressing alternative splicing are still presence. As both of them are still challenging problems and short board of all gene prediction programs, we will try to further exploit it individually in future work.

2 Method

2.1 Algorithmic Issues of the GHMM

Hidden Markov Models (HMM) has been used in pattern recognition for decades and its applicability to computational biology has also been widely recognized. But as we know, a standard Hidden Markov model is just a state-based generative model which transitions stochastically from state to state and emits a single symbol from each state [8]. Although it can produce a certain effect in gene prediction, the recognition accuracy is still far from satisfactory. The GHMM have a better performance by allowing an individual state to emit a string of symbols rather than only one symbol at a time. The model is generally parameterized by its transition probabilities, state duration (i.e., feature length) probabilities, and state emission probabilities. These probabilities influence the output of the model by determining which sequences are more likely to be emitted and which series of states are more likely to be visited by it.

Eukaryotic gene prediction with a GHMM means to decode an input sequence into a most probable set of putative functional segments having a specific biological significance [9]. Suppose that X denotes an input DNA sequence with a length *n*, $\mathrm{x_i}(1 \le \mathrm{i} \ll \mathrm{k})$ denotes a subsequence of X,and its length is $\mathrm{d_i}(1 \le \mathrm{d_i} \le \mathrm{n})$, we can get that $\mathrm{X} = \mathrm{x_1 x_2 \cdots x_k}$ (the concatenation of subsequences), and define Ø is a correct parse corresponding to the input sequence, having that $\emptyset = \{(\mathrm{q_1}, \mathrm{x_1}), \cdots, (\mathrm{q_i}, \mathrm{x_i}), \cdots (\mathrm{q_k}, \mathrm{x_k})\}(1 \le \mathrm{i} \le \mathrm{k})$, and $\mathrm{q_i}$ denotes a hidden state which signifies a specific functional segment mentioned above. But in general, we still need to supplement two additional states producing no output, as start and end flags of decoding operation of a program. And then, how to set the optimal value of Ø is what we concern and difficult to gain. In the case of standard Hidden Markov Models, the well-known Viterbi algorithm [10], a dynamic programming algorithm with running time linear to the sequence length for a fixed number of states, is the most classic means to solve with this problem, similarly, it is also applicable to the case of GHMM. However, since each state can emit more than one symbol at a time, the algorithm needs to be modified to result in the following optimization problem [11]:

$$\begin{aligned}
\Phi_{optimal} &= \arg\max p(\Phi|X) \\
&= \arg\max \frac{p(\Phi, X)}{p(X)} \\
&\simeq \arg\max p(\Phi, X) \\
&= \arg\max p(X|\Phi)p(\Phi) \\
&= \arg\max \coprod_{i=1}^{k} p_e(x_i|q_i, d_i)p_t(q_i|q_{i-1})p_d(d_i|q_i)
\end{aligned} \tag{1}$$

where $\mathrm{P_e}(\mathrm{x_i}|\mathrm{q_i}, \mathrm{d_i})$ means the probability that state $\mathrm{q_i}$ emits the subsequence $\mathrm{x_i}$, given duration $\mathrm{d_i}, \mathrm{P_t}(\mathrm{q_i}|\mathrm{q_{i-1}})$ denotes the probability that the GHMM translates from $\mathrm{q_{i-1}}$ state to state $\mathrm{q_i}$; and $\mathrm{P_d}(\mathrm{d_i}|\mathrm{q_i})$ is the probability that state $\mathrm{q_i}$ has the duration $\mathrm{d_i}$, the arg max is to select the best one from all parses of the DNA sequence into well-formed exon-intron structures.

We introduce a common approach, named the Prefix Sum Arrays (PSA), to evaluate Eq. 1. According to a dynamic programming algorithm, the method needs to allocate several arrays for one per variable-length feature state and assess them left-to-right along the length of the input sequence. It can also conclude that the values in the aforementioned arrays represent cumulative scores for prefixes of the sequence only in term of the surface meaning of its name. Here, we show its recursive expressions of the GHMM in log space as follows:

$$\begin{aligned}
&R_I(q_j, r_j) = \arg\max_{q_i}(R_I(q_i, r_i) + R_T(q_i, q_j) + R_D(q_i, q_j) + R_C(q_i, q_j, r_j)) \\
&q_i, q_j \in Q
\end{aligned} \tag{2}$$

In Eq. 2, Q denotes the set of states in GHMM, $R_I(q_i, r_i)$ denotes the logarithmic inductive score for signal q_i in phase r_i, and the next three expressions respectively mean the logarithmic scores of state translation from q_i to q_j, state duration of content region delimited by signals q_i and q_j, and sequence emission between current signal q_j and predecessor q_i in phase r_i, additionally, it is still necessary to emphasize that $r_j = r_i$ or $r_j = (r_i \pm \Delta)$ mod 3 (Δ denotes the sequence length of special putative state), depending on the different situations.

2.2 Modeling Gene Structure

To expound completely the process of gene prediction based on a GHMM, Fig. 1 shows the states of the Hidden Markov Models in the system and some certain probabilities of possible transitions between them (as to others that cannot be depicted explicitly at the arrows, their values are always 1). In Fig. 1, Esng denotes a single exon gene; EI, E and EF respectively denote the first, internal and last exon of a multi exon gene (the exon only referred to the coding part of exons); I is the intron, IR is the intergenic region between genes, DSS and ASS separately are the donor and acceptor splice sites including branch point, as for the states S and T, they are the start codon emitting the string ATG with probability 1 and stop codon generally only including TAG, TGA and TAA whose emission probabilities are respectively 24 %, 48 % and 28 %. Furthermore, the states with names beginning with r mean to be on the reverse strand, and the exponents (0, 1, 2) stand for the phase of the reading frame, and for an exon it denotes the position of the last coding nucleotide of the exon in its codon.

In the GHMM, each state emits a random DNA string with random length, and their emission probabilities mainly depends on the annotated sequences which correspond to the respective biometrics in training set. In order to seize the feature information of this distribution for each state, we mainly made use of five established models, a Markov chain, a higher order windowed weight array model (WWAM), a weight array model (WAM) [12], simple interpolated Markov Models (IMM) [13] and the method based on similarity weighting of sequence patterns [6], whose good results have been verified by other gene finders.

In term of the details of the Markov Chain, WAM and WWAM, there is no need to elaborate too much again, since that they have been widely used in bioinformatics for many years, and here, we briefly illustrate our usage on them. We adopt a Markov model of order 5 to the model of non-coding region such as I, rI and IR as mentioned above meanwhile using a WAM of order 2 and a WWAM of order 2 and of window size 5 in other related states. As for the IMM, it's a special case in our coding models, in which only the transition probabilities of order 5 and 4 are considered and the respective interpolation weights are either 0 or 1 with the frequency threshold of occurrence of the given string in training set 400. Finally, we focus on the method of similarity-based weighting of sequence patterns, which is solely applied in the DSS model. Given a fixed sequence pattern size, training patterns $\mathrm{q}_1, \mathrm{q}_2 \cdots, \mathrm{q}_\mathrm{m}$ and a similarity scoring function s, weighting pairs of patterns, we estimate the probability that a random pattern equals a given pattern q as

$$\mathrm{p(q)} = \mathrm{c} \sum_{\mathrm{i}=1}^{\mathrm{m}} \mathrm{s(q, q_i)} \tag{3}$$

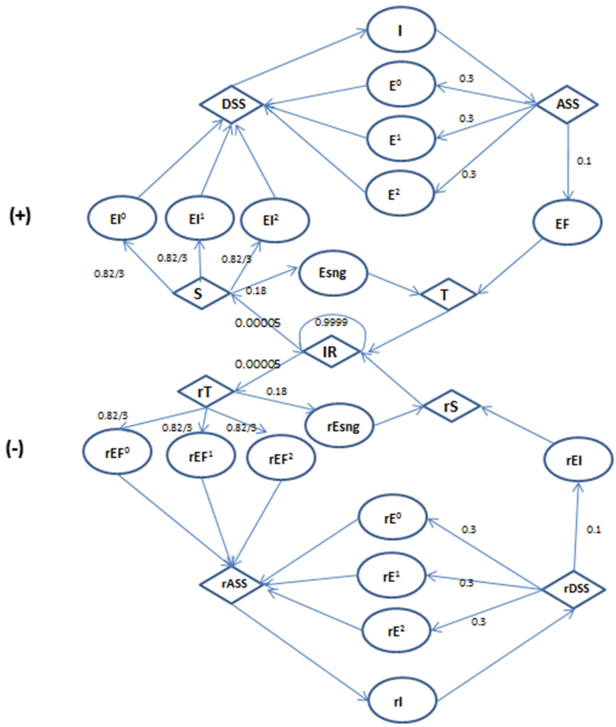

Fig. 1. The GHMM topology of our system

where c is a modulus keeping that the sum of all p(q) is 1. Regarding the similarity scoring function s, we follow the definition in AUGUSTUS as

$$s(r,q) = \begin{cases} 1 & \text{if } r{=}q \\ 0.001 & \text{if } r \text{ and } q \text{ differ at exactly one pos}, \\ 0 & \text{otherwise} \end{cases} \tag{4}$$

and the resulting distribution obtained by this way is the discretely smoothed empirical distribution which respects the complicated statistical dependencies that exist between the nucleotide positions.

Figure 2 shows detailed model distribution of human gene structure with single and multiple exons in our system. According to it, we once more simply describe the emission distribution for those states which are not mentioned above. The models of fixed length, translation initiation motif and ASS model respectively emitting 20 and 23 nucleotides per time, are trained by the WWAM of order 2 and window size 5, while the model of translation end motif emitting 30 nucleotides per time introduces the WAM of order 2 to evaluate.

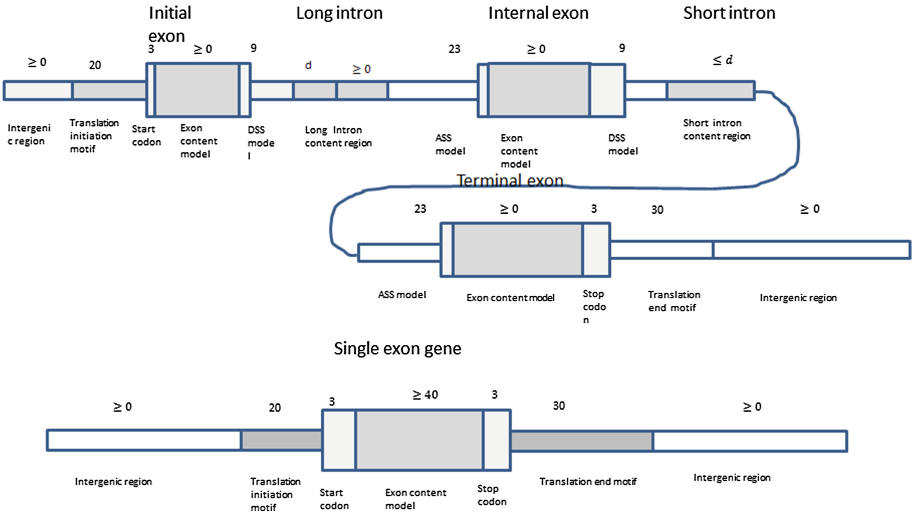

Fig. 2. My model distribution of gene structure with single and multiple exons

2.3 Intron Length Model

In the GHMM, length distribution of states with variable length, such as intron and exon, is also significant information which can determine the prediction accuracy. We adopt a typical smoothing technique using a kernel estimator with discrete Gaussian kernel function and variable bandwidth to evaluate them. And because of compact length density of coding exons (only 0.3 % of the human exons were longer than 3000 nucleotides), the evaluation effect is fairly satisfactory. However, to a long biological intron, for example, the human neurexin-3 gene on chromosome 14 has an intron of length 479 Kb, results in a large range of length span, and it is practically infeasible to explicitly model the whole length distribution in a HMM. In order to deal with this problem, we have combined the method mentioned above with a simple geometric distribution to model it. Define a length threshold d, a probability parameter p for determining to choose a short intron, and parameter q which is only used by the geometric distribution function, the concrete formula expression as follows:

$$P(M = l) = \begin{cases} pP(L = l)/P(L \le d) & l \le d \\ (1 - p)(1 - q)^{l-d-1}q & l > d \end{cases} \tag{5}$$

where M denotes a model variable to be evaluated and L denotes the variable based on the discrete Gaussian kernel function and variable bandwidth. Firstly, to keep the continuity of functions, we can get an equation that $P_{-}(M = d + 1) = P_{+}(M = d)$, stating that there is no jump in the distribution of M between positions d and $d + 1$. Secondly, we need to set q so that the expectations of M and L are equivalent, when $M > d, L > d$, for instance, $d + 1/q = E[L|L > d]$. Thirdly, to better take into account both of accuracy (large d) and speed (small d), we choose the smallest d so that parameter p approximately equal to $P(L \leq d)$. Finally, combining all of the three points, we have obtained a group of evaluation values using my training set that $q \approx 1/1764, p \approx 0.47, d = 592$.

2.4 Training Data Set

The training set was retrieved from 22 autosomal sequences of human genome, which is in the light of the corresponding version of the gene annotation files issued in Genbank. After getting rid of the case in which sequences are overlapping with another sequence by our self-made checking procedure, we luckily extract most of gene sequences with single transcription as the master training sets, about 941 sequences, and randomly select a certain number of genes with multiple transcriptions, e.g. approximately 400, as the addition of our training set (naturally, the genes of test sets provided by AUGUSTUS, h178 and sag178, have been removed from it). Then, making most use of the two data sets, we train the each relevant state model in our GHMM.

3 Results and Discussion

We tested our program on two data sets, called h178 and sag178, which can be downloaded from the official website of AUGUSTUS. The h178 is a set of 178 human genomic sequences which are from EMBL and have been used by the author of GENSCAN for evaluation; each sequence only contains one complete gene and their mean sequence length is 7169 bases, shortest 622 and longest 86640 bases. The sag178 has the same 178 human genes in which 40 genes are single exon genes, but all of them are included in a set of 43 sequences on both strands. These sequences are taken from Guigo et al. (2000) as like the h178 and have been done some necessary special process, their mean length is 177 kilo bases (shortest 70, longest 282) and the average number of genes is 4.1.

Table 1. Accuracy results on human data sets h178

		base	exon	gene
GENEID	sn	89	66	14
	sp	91	75	13
GENSCAN	sn	97	83	40
	sp	86	75	36
AUGUSTUS	sn	93	80	46
	sp	90	80	45
OUR SYSTEM	sn	92	74	31
	sp	89	72	27

In order to evaluate the gene prediction performance, we also adopted the usual measures, sensitivity and specificity, for a feature such as base, exon and gene. The sensitivity is defined as the number of correctly predicted features divided by the number of annotated features. The specificity is the ratio of the number of correctly predicted features to the number of predicted features. A predicted exon is considered to be correct if both splice sites are at the annotated position of an exon. A gene is considered to be predicted correctly if all the exons are correctly predicted and no additional exons are not in the annotation. Predicted partial genes were counted as predicted genes. The testing results on both of test sets are depicted in the following tables (Tables 1 and 2).

Table 2. Accuracy results on human data set sag178

		base	exon	gene
GENEID	sn	89	67	17
	sp	78	60	17
GENSCAN	sn	94	68	18
	sp	64	45	14
AUGUSTUS	sn	93	78	40
	sp	81	71	35
OUR SYSTEM	sn	90	73	23
	sp	76	61	19

Comparing the above two tables carefully, we can analyze that my system can have similar prediction accuracy with AUGUSTUS and GENSCAN in term of the mean of sensitivity and specificity on the base and exon level. GENSCAN is more sensitive, AUGUSTUS is more specific, and our program is indeed worse a little in both aspects but superior than GENEID. Whether on long or short gene set, our model predicted exactly more genes than GENEID and GENESCAN, stating that the design combining main model structures of the AUGUSTUS and GENSCAN with the more precise evaluation of length distribution in intron is effective. However, the number of genes predicted correctly is only slightly higher than GENSCAN's while far lower than AUGUSUTS', we guess, which is likely due to ignoring the influence of the GC-content in genes. Besides, comparatively speaking, our model tends to produce many more genes in which only partly exons are evaluated correctly and is therefore less specific than others. To solve with this problem, it is necessary to introduce some further follow-up design and optimization.

4 Conclusion

In our paper, with the integration and modification of some mainly related submodels, we personally implemented a GHMM-based gene prediction system. Despite a certain degree of performance promotion, there is still a lot of space to further improve by considering the influence of the GC-content in genes and training a better classification model of spite sites with other superior machine learning methods like SVM. In the future, we will continue to deepen our research from two above-mentioned aspects.

Acknowledgement. This article is partly supported by the Shenzhen Municipal Science and Technology Innovation Council (Nos. JCYJ20140904154645958).

References

1. Cairui, L., Changsong, Z., Guoli, S.: Recent progress in gene mapping through high-throughput sequencing technology and forward genetic approaches. Yi chuan = Hereditas/ Zhongguo yi chuan xue hui bian ji **37**(8), 765–776 (2015)
2. Burge, C., Karlin, S.: Prediction of complete gene structures in human genomic DNA. J. Mol. Biol. **268**(1), 78–94 (1997)
3. Burset, M., Seledtsov, I.A., Solovyev, V.V.: Analysis of canonical and non-canonical splice sites in mammalian genomes. Nucleic Acids Res. **28**(21), 4364–4375 (2000)
4. Guigó, R., et al.: Prediction of gene structure ☆. J. Mol. Biol. **226**(1), 141–157 (1992)
5. Haussler, D., David, K., Reese, M.G., Eeckman, F.H.: A generalized hidden Markov model for the recognition of human genes in DNA. In: Proceedings of the International Conference on Intelligent Systems for Molecular Biology, St. Louis (1996)
6. Stanke, M., Waack, S.: Gene prediction with a hidden Markov model and a new intron submodel. Bioinformatics **19**(suppl 2), 215–225 (2003)
7. Fickett, J.W.: Finding genes by computer: the state of the art. Trends Genet. **12**(8), 316–320 (1996)
8. Krogh, A., Mian, I.S., Haussler, D.: A hidden Markov model that finds genes in E. coli DNA. Nucleic Acids Res. **22**(22), 4768–4778 (1994)
9. Salzberg, Steven L., D. B. Searls, and S. Kasif. “Computational methods in molecular biology.” Computational Methods in Molecular Biology49.2(1999):191-192
10. Ryan, M.S., Nudd, G.R.: The viterbi algorithm. Warwick Res. Rep. Rr **37**(2), 160–163 (1993)
11. Majoros, W.H., et al.: Efficient decoding algorithms for generalized hidden Markov model gene finders. BMC Bioinform. **6**(2), 8–16 (2005)
12. Zhang, M.Q., Marr, T.G.: A weight array method for splicing signal analysis. Comput. Appl. Biosci. Cabios **9**(5), 499–509 (1993)
13. Salzberg, S.L., et al.: Microbial gene identification using interpolated Markov models. Nucleic Acids Res. **26**(2), 544–548 (1998)