Improving the Prediction Accuracy of Gene Structures in Eukaryotic DNA with Low C+G Contents

Yanhong Zhou[1], Huili Zhang[1], Lei Yang[1], and Honghui Wan[2]

[1] Hubei Bioinformatics and Molecular Imaging Key Laboratory, Huazhong
University of Science and Technology, Wuhan, Hubei 430074, China
[2] Laboratory of Bioinformatics, Maryland Institute of Dynamic Genomics,
Silver Spring, MD 20906, USA

yhzhou@hust.edu.cn

Abstract

We have developed a gene prediction program GeneKey. When trained with the widely used dataset collected by Kulp and Reese, GeneKey can achieve high prediction accuracy for genes with moderate and high C+G contents. However, the prediction accuracy is much lower for CG-poor genes. To tackle this problem, we construct a new LCG316 dataset composed of gene sequences with low C+G contents. For CG-poor genes, the prediction accuracy of GeneKey when trained with LCG316 dataset has been improved prominently. Further statistical analysis demonstrates that some structure features, such as splicing signals and codon usage, of CG-poor genes are quite different from that of CG-rich ones. The combination of the two datasets enables GeneKey to get high and balanced prediction accuracy for both CG-rich and CG-poor genes. The results of this work imply that careful construction of training dataset is very important for improving the performance of various prediction tasks. The GeneKey program is available at http://infosci.hust.edu.cn.

Keywords: DNA sequence, prediction of gene structure, prediction of protein coding region

I. Introduction

Accurate computational identification of eukaryotic gene structures is still a challenging problem in bioinformatics. This problem has attracted extensive researches and various approaches and a number of computational tools for eukaryotic gene identification have been developed [1-13]. Some of these tools have been widely used to identify putative genes in uncharacterized DNA, and have played a significant role in the genome annotation of human and other model organisms [14, 15].

With more and more gene-finding programs becoming available, the evaluation of such programs has also been reported for several times [16, 17]. These evaluations suggest that the accuracy of gene-finding programs is inclined to be dependent on the C+G contents of DNA sequences. According to the systematic evaluation made by Rogic *et al*. [17] with a carefully selected test set, the HMR195, the prediction accuracy of some well-known gene-finding programs is significantly lower for DNA sequences with low C+G contents. However, as human and other mammalian genomes [14, 15, 18] contain a considerable part of DNA sequences with low C+G contents, it is important to accurately identify the locations and structures of CG-poor genes.

A gene prediction program is composed of two parts: a computational model and a training dataset for determining the parameters of the model. Various computational models such as the Hidden

Markov Model (HMM) have been used in gene prediction programs. A widely used training dataset is the set of gene sequences collected by Kulp and Reese, which have been used to determine the parameters of many gene prediction models. We have developed a eukaryotic gene prediction program, called GeneKey (http://infosci.hust.edu.cn), in which we used a multistage optimization model configuration to process various kinds of information and to deal with different sub-problems associated with the identification of genes and protein-coding regions [12]. When trained with the Kulp-Reese dataset, GeneKey can get higher prediction accuracy than other programs for DNA sequences with C+G content >40%, however, the prediction accuracy is much lower for DNA sequences with C+G content <40%.

In this study, in order to improve the prediction accuracy for CG-poor genes, we construct a new LCG316 dataset composed of 316 human gene sequences with low C+G contents. Both the LCG316 dataset and the Kulp-Reese dataset are used to train our gene prediction program GeneKey, and the independent dataset HMR195 is used to test the prediction accuracy. The results demonstrate that, for CG-poor genes, the prediction accuracy of GeneKey when trained with the LCG316 dataset is much higher than that when trained with the Kulp-Reese dataset. To make clear the course of this improved performance, further statistical analysis is carried out, and it is found that some features of gene structure, such as splicing signals and codon usage, obtained from the LCG316 dataset are quite different from that obtained from the Kulp-Reese dataset, means that the LCG316 dataset can reflect the structural features of CG-poor genes better than the Kulp-Reese dataset. The combination of the two datasets enables GeneKey to get high and balanced prediction accuracy for both CG-rich and CG-poor genes.

II. Methods

A. *Sequence Datasets*

We employ the HMR195 dataset (http://www.cs.ubc.ca/-rogic/evaluation/) as the test set in this work. This dataset, which consists of 195 human and rat gene sequences, was constructed specifically for the evaluation of gene prediction programs [17].

Two training datasets are used in this study. One is the Kulp-Reese dataset and the other is the LCG316 dataset constructed in this study. The Kulp-Reese dataset (http://www.fruitfly.org/seq_tools/datasets/Human/), which consists of 462 non-redundant multi-exon genes, is a benchmark data set and has been widely used to train many powerful gene prediction algorithms. The LCG316 dataset comprises 316 human gene sequences with C+G content < 45% and can be divided into two parts. One consists of 98 CG-poor gene sequences taken from the Kulp-Reese dataset. The other contains 218 carefully selected CG-poor human gene sequences, which have no significant sequence similarity between each other and with the gene sequences in both the HMR195 and the Kulp-Reese dataset.

Note that each sequence in the training and test datasets contains a single gene without alternatively spliced forms. Some additional consistency constraints are also enforced, e.g. there should be no in-frame stop codons in the annotated coding regions, and the splicing signals should match the minimal consensus (GT for donor splicing sites and AG for acceptor splicing sites).

B. *The GeneKey Program*

We have developed a program, called GeneKey, for the prediction of protein coding regions and gene structures in eukaryotic DNA. A four-stage optimization model configuration is adopted in GeneKey to process various kinds of information and deal with different sub-problems associated with the identification of genes and protein-coding regions [12].

The first stage is called "feature modelling". Four statistical features are currently used in the Genekey, including the signals of functional sites (donor splicing sites, acceptor splicing sites, translation initiation sites and termination sites), codon usage preference, length distributions of exon regions, and the correlation between the C+G content of exons and the adjacent introns. Different models are used for identifying these features in GeneKey. For example, the WMM (Weight Matrix Model) is used to recognize the translation initiation and termination sites; the WAM (Weight Array Model) is employed to detect acceptor splicing sites and donor splicing sites; the in-frame hexamer usage model is applied to score the coding potential of a DNA fragment according to the di-codon usage.

The second stage is called "unit modelling". Six basic units of the gene structure are currently considered in the GeneKey, including the single exon, internal exon, initiation exon, termination exon, intron and intergenic region. LDA (Linear discriminant analysis) method is used to model these basic units by combining related feature models created in the first stage. For example, the internal exon model is the weighted linear sum of such feature models as acceptor splicing site, donor splicing site, codon usage, length distribution of internal exons, and the C+G content distribution, where the weight coefficients are determined by the LDA method.

The third stage is called "gene modelling", including the modelling of single exon genes and multi-exon genes. The model of single exon gene is the same as that of single exon built in the second stage. The model of multi-exon gene, however, is the weighted linear sum of such unit models as the initiation exon, internal exon, termination exon and intron, where the weight coefficients are also determined by the LDA method.

The last stage is called "genome modelling", which aims at dealing with long DNA sequences containing many genes. The genome model is the weighted linear sum of such gene and element models as the single-exon gene, multi-exon gene and intergenic region, where the weight coefficients are also calculated by the LDA method.

Given a training dataset, all the parameters of above models are determined stage by stage. The genome model is then used as the objective function to predict the most possible gene structures in uncharacterized DNA sequences, which is implemented with the dynamic programming approach.

C. *Experiments and Performance Evaluation*

We use the Kulp-Reese dataset and the LCG316 dataset to train the models of GeneKey respectively, and then use the HMR195 dataset to test the prediction performance of GeneKey. The performance is evaluated at both the nucleotide level and exon level.

At the nucleotide level, let *TP* be the number of correctly predicted coding nucleotides, *TN* be the number of correctly predicted non-coding nucleotides, *FP* be the number of incorrectly predicted coding nucleotides, and *FN* be the number of incorrectly predicted non-coding nucleotides. Then the prediction accuracy is measured by the sensitivity $Sn = TP/(TP+FN)$, specificity $Sp = TP/(TP+FP)$, and the approximate correlation *AC* defined by

$$AC = \frac{1}{2} \cdot \left(\frac{TP}{TP+FN} + \frac{TP}{TP+FP} + \frac{TN}{TN+FP} + \frac{TN}{TN+FN} \right) - 1 \cdot$$

At the exon level, let *ETP* be the number of predicted exons which are identical to the corresponding real exons, *EFP* be the number of predicted exons which are not real exons, *EFN* be the number of real exons that are not correctly predicted. Then the prediction accuracy is measured by sensitivity *ESn* = *ETP*/(*ETP*+*EFN*) and specificity *ESp* = *ETP*/(*ETP*+*EFP*).

III. Results

When trained with the Kulp-Reese dataset and tested with the HMR195 dataset, the prediction accuracy of GeneKey for genes with different C+G contents, which are measured by the approximate correlation *AC* at the nucleotide level and (*ESn*+*ESp*)/2 at the exon level, is given in Table 1. Also shown in this table are the accuracies of 7 famous gene prediction programs (FGENES, GeneMark.hmm, Genie, Genscan, HMMgene, Morgan, and MZEF), which are evaluated by Rogic *et al*. with the same test dataset [17]. The HMR195 dataset is partitioned into 4 groups according to the C+G contents of gene sequences. Group 1 consists of 14 genes with C+G content < 40%; Group 2 consists of 69 genes with C+G contents ranging from 40% to 50%; Group 3 consists of 93 genes with C+G contents ranging from 50% to 60%; and Group 4 consists of 19 genes with C+G content ≥ 60%. For each group, *AC* and (*ESn*+*ESp*)/2 are averaged over all gene sequences in it.

Table 1. Prediction accuracy of GeneKey (trained with the Kulp-Reese dataset) and other 7 programs

Program	C+G < 40%		40% ≤ C+G < 50%		50% ≤ C+G < 60%		CG ≥ 60%	
	AC	(*ESn*+*ESp*)/2	AC	(*ESn*+*ESp*)/2	*AC*	(*ESn*+*ESp*)/2	AC	(*ESn*+*ESp*)/2
GeneKey	0.85	0.64	0.92	0.74	0.91	0.79	0.94	0.79
FGENES	0.84	0.70	0.81	0.64	0.85	0.71	0.87	0.66
GeneMark	0.79	0.48	0.80	0.46	0.87	0.62	0.85	0.48
Genie	0.85	0.69	0.85	0.60	0.92	0.75	0.87	0.79
Genscan	0.94	0.80	0.91	0.66	0.91	0.74	0.88	0.70
HMMgene	0.91	0.76	0.90	0.73	0.92	0.79	0.91	0.77
Morgan	0.65	0.29	0.72	0.49	0.69	0.43	0.69	0.37
MZEF	0.66	0.71	0.65	0.50	0.70	0.62	0.58	0.53

It can be seen from Table 1 that the overall performance of GeneKey when trained with the Kulp-Reese dataset is comparable to other 7 well known gene prediction programs. Particularly, for those genes with C+G content ≥ 40%, the prediction accuracy of GeneKey is better than other programs at both the nucleotide level and exon level, demonstrating the efficiency of the multilevel optimization models used in Genekey. For genes with C+G content < 40%, however, the prediction accuracy of GeneKey is much lower than that for genes with C+G content ≥ 40%, also lower than that of some gene finding programs such as Genscan and HMMgene.

As shown in Table 2, for genes with C+G content < 40%, GeneKey achieves a much higher prediction accuracy when trained with the LCG316 dataset than with the Kulp-Reese dataset. The sensitivity/specificity is improved to 0.97/0.96 from 0.80/0.93 at the nucleotide level, and jumps to 0.85/0.88 from 0.53/0.76 at the exon level. Compared with the results of other 7 programs, GeneKey has the best prediction performance. A dramatic improvement is made for the exon level prediction accuracy in terms of (*ESn*+*ESp*)/2 (0.86 versus 0.80 for the best program Genscan), and a small improvement is also made for the nucleotide level prediction accuracy in terms of the approximate correlation *AC* (0.96 versus 0.94 for Genscan). Combining Table 1 and 2, we can see if the GeneKey is trained with the LCG316 dataset to predict genes with C+G content < 40% and trained with the Kulp-Reese dataset to predict genes with C+G content ≥ 40%, then, GeneKey has better performance at both the nucleotide and exon levels than the other 7 gene finding programs for all 4 groups of gene sequences in the HMR195 dataset.

Table 2. Prediction accuracy of GeneKey for genes with C+G content lower than 40%, when trained with two different datasets

Training dataset	Nucleotide level			Exon level		
	Sn	*Sp*	*AC*	*ESn*	*ESp*	(*ESn*+*ESp*)/2
Kulp-Reese set	0.80	0.93	0.85	0.53	0.76	0.64
LCG316 set	0.97	0.96	0.96	0.85	0.88	0.86

IV. Discussion

Most of the previous studies on eukaryotic gene prediction have focused on the improvement of modeling and computational techniques rather than on the improvement of training datasets. In this work, we have improved the prediction accuracy of gene finding programs by constructing suitable training datasets. Some of gene prediction programs, including Genie, HMMgene, and GeneKey, trained with the Kulp-Reese dataset usually get lower prediction accuracy for CG-poor genes than those genes with average or high C+G content. So there may be a significant distinction between the structural features of CG-poor genes and that of genes with average or high C+G contents, and the Kulp-Reese set may not sufficiently reveal the structural characteristics of CG-poor genes. That is, unsuitable selection of training set is probably the main cause of low prediction accuracy for gene sequences with low C+G contents, although other factors such as modeling techniques may also have some influences. These ideas are confirmed by our experimental results. When GeneKey is trained with the LCG316 dataset, a new dataset composed of 316 human gene sequences with low C+G content, and tested with the same HMR195 dataset, the prediction accuracy for CG-poor gene sequences are much higher than that when trained with the Kulp-Reese dataset. The underlying cause for the improved accuracy is that the LCG316 dataset can reflect the structural features of CG-poor genes better than the Kulp-Reese dataset, which can be seen from the following analysis.

A. *Codon Usage*

Codon usage is one of the most important features used for the prediction of protein coding regions and gene structures. Codon usage frequencies are different in the coding regions of gene sequences with different C+G contents. One of the important reasons that Genscan has substantially higher accuracy than other gene prediction programs for CG-poor genes is that a subset of 638 cDNAs with low C+G contents has been used for training the program.

We compared the trinucleotide statistics based on the Kulp-Reese dataset and the LCG316 dataset. The results are given in Table 3, which demonstrate that the codon usage frequency distribution for the Kulp-Reese dataset is extremely different from that for the LCG316 dataset. Furthermore, the C+G content distribution at three codon positions is also analyzed. The C+G contents of the first, second, and third codon position in the coding regions of genes in the Kulp-Reese set are 59.3%, 43.6 %, and 67.1%, respectively. However, in the coding regions of genes in the LCG316 set, the C+G contents of the first, second, and third codon position are 52.6%, 39.3%, and 48.0%, respectively. The average C+G content of the coding regions of genes in the Kulp-Reese set is 63.3%, while the average C+G content of the coding regions of genes in the LCG316 dataset is only 46.6%. Note that the in-frame hexamer usage model is used in GeneKey to estimate the likelihood of a DNA sequence coding for a protein in terms of codon usage preference. The higher C+G contents in the coding regions of genes in the Kulp-Reese dataset makes GeneKey, when trained with the Kulp-Reese dataset, get a lower prediction accuracy for CG-poor genes.

Table 3. Codon usage in Kulp-Reese dataset /LCG316 dataset (per 1000 codons)

Codon	Frequency	Codon	Frequency	Codon	Frequency	Codon	Frequency
AAA	17.68/35.35	AAG	35.53/31.48	GGC	30.69/16.21	GTG	32.96/23.22
ACA	11.23/17.76	ACG	7.15/4.56	GTC	16.20/12.58	TAG	0.00/0.00
AGA	8.01/15.18	AGG	11.02/10.96	TAC	18.30/13.72	TCG	5.07/2.70
ATA	4.43/10.41	ATG	21.75/23.38	TCC	18.42/14.55	TGG	13.69/10.60
CAA	9.49/17.89	CAG	36.16/30.83	TGC	14.19/9.62	TTG	10.27/15.66
CCA	13.87/18.16	CCG	8.20/3.96	TTC	24.42/17.27	CCT	17.52/18.08
CGA	5.93/6.82	AAC	20.92/19.71	AAT	12.76/25.18	CGT	5.39/4.48
CTA	5.43/8.22	ACC	22.58/14.99	ACT	10.90/16.16	CTT	9.71/16.23
GAA	22.82/41.28	AGC	20.45/15.50	AGT	8.14/15.15	GAT	17.95/28.57
GCA	13.20/17.99	ATC	23.95/18.37	ATT	12.42/21.30	GCT	19.28/20.50
GGA	14.37/20.9	CAC	15.47/12.35	CAT	8.08/14.10	GGT	12.50/11.94
GTA	5.15/9.24	CCC	24.88/12.49	CGG	12.72/6.68	GTT	8.19/15.09
TAA	0.00/0.00	CGC	16.14/5.74	CTG	50.76/27.42	TAT	10.20/15.86
TCA	7.81/14.46	CTC	22.04/14.98	GAG	45.38/31.56	TCT	11.14/18.43
TGA	0.00/0.00	GAC	28.52/22.35	GCG	9.78/4.58	TGT	8.12/12.24
TTA	3.27/10.66	GCC	35.80/19.54	GGG	17.86/12.26	TTT	13.50/22.26

To access the influence of the coden usage difference given in Table 3, the in-frame hexamer usage model used in GeneKey is trained with the LCG316 and Kulp-Reese set separately, and tested on Group 1 sequences of the HMR195 dataset. This test set is composed of positive samples and negative samples. The positive samples include all the internal exons, while all the pseudo-exon sequences extracted from intron regions are considered as the negative samples. The testing results demonstrate that the in-frame hexamer usage model performs much better in the recognition of coding regions of CG-poor genes when trained with the LCG316 set (Data not shown).

B. *Splicing Signals*

The acceptor and donor splicing signals are the most critical information for the prediction of exact exons. We have analyzed splicing signals of genes with the LCG316 and the Kulp-Reese datasets, respectively. The results show that there is a strong similarity in donor splicing site profiles except for the third downstream position (as shown in Table 4), but a significant difference in the acceptor splicing site profiles between the two datasets (as shown in Table 5). There is a CT-rich region upstream the acceptor splicing sites in the genes of the Kulp-Reese dataset. In this region, C+T content is much higher than A+G content, and the frequencies of C and T are almost the same. In the genes of LCG316 dataset, however, the frequency of T is much higher than the 3 other nucleotides.

To access the influence of the splicing signal difference given in Table 5, the WAM acceptor model used in the GeneKey is trained with the LCG316 dataset and the Kulp-Reese dataset respectively, and the performances are evaluated on Group 1 sequences of the HMR195 dataset. The results demonstrate that the WAM acceptor model achieves a better performance trained with LCG316 dataset than with the Kulp-Reese dataset (data not shown), which contributes to the improved prediction accuracy of GeneKey for CG-poor genes, especially the exon level accuracy.

V. Conclusion

We have developed a gene prediction program GeneKey, which can achieve high prediction accuracy for genes with moderate and high C+G contents, but the prediction accuracy of which is

Table 4. Donor splicing site profiles of Kulp-Reese dataset /LCG316 dataset

Position	P(A)	P(C)	P(G)	P(T)
-3	0.33/0.36	0.36/0.33	0.19/0.18	0.12/0.13
-2	0.59/0.64	0.14/0.10	0.13/0.11	0.14/0.15
-1	0.09/0.13	0.03/0.03	0.79/0.77	0.08/0.08
1	0.00/0.00	1.00/1.00	0.00/0.00	0.00/0.00
2	0.00/0.00	0.00/0.00	0.00/0.00	1.00/1.00
3	0.49/0.72	0.03/0.02	0.46/0.23	0.02/0.03
4	0.71/0.73	0.08/0.05	0.12/0.09	0.09/0.14
5	0.07/0.10	0.05/0.04	0.84/0.77	0.05/0.08
6	0.15/0.19	0.16/0.12	0.22/0.14	0.47/0.55

Table 5. Acceptor splicing site profiles of Kulp-Reese dataset /LCG316 dataset

Position	P(A)	P(C)	P(G)	P(T)
-25	0.22/0.29	0.30/0.18	0.16/0.13	0.32/0.41
-24	0.20/0.28	0.32/0.18	0.17/0.14	0.32/0.40
-23	0.22/0.27	0.30/0.17	0.17/0.14	0.31/0.42
-22	0.21/0.26	0.32/0.19	0.17/0.14	0.31/0.42
-21	0.19/0.24	0.33/0.19	0.16/0.13	0.32/0.44
-20	0.18/0.23	0.32/0.19	0.16/0.13	0.34/0.45
-19	0.16/0.20	0.33/0.19	0.16/0.13	0.35/0.48
-18	0.14/0.18	0.34/0.21	0.15/0.12	0.36/0.49
-17	0.13/0.17	0.33/0.19	0.17/0.13	0.38/0.51
-16	0.13/0.16	0.35/0.20	0.14/0.11	0.38/0.53
-15	0.11/0.14	0.35/0.21	0.12/0.11	0.42/0.53
-14	0.09/0.12	0.37/0.20	0.13/0.11	0.42/0.57
-13	0.09/0.11	0.35/0.21	0.12/0.10	0.45/0.58
-12	0.08/0.10	0.36/0.21	0.11/0.10	0.45/0.59
-11	0.08/0.11	0.33/0.19	0.11/0.08	0.48/0.62
-10	0.07/0.10	0.37/0.21	0.11/0.09	0.46/0.60
-9	0.07/0.12	0.39/0.23	0.12/0.11	0.42/0.55
-8	0.09/0.13	0.41/0.26	0.12/0.09	0.38/0.52
-7	0.08/0.14	0.42/0.25	0.09/0.08	0.41/0.53
-6	0.07/0.11	0.45/0.25	0.07/0.05	0.41/0.59
-5	0.07/0.11	0.39/0.21	0.06/0.05	0.48/0.63
-4	0.22/0.27	0.34/0.23	0.22/0.16	0.22/0.34
-3	0.04/0.08	0.74/0.56	0.00/0.02	0.22/0.34
-2	1.00/1.00	0.00/0.00	0.00/0.00	0.00/0.00
-1	0.00/0.00	0.00/0.00	1.00/1.00	0.00/0.00
1	0.23/0.28	0.14/0.13	0.53/0.46	0.10/0.13
2	0.22/0.28	0.20/0.16	0.25/0.17	0.32/0.39
3	0.24/0.30	0.25/0.19	0.25/0.21	0.26/0.31

much lower for CG-poor genes when trained with the widely used dataset collected by Kulp and Reese. To tackle this problem, we constructed a new LCG316 dataset composed of gene sequences with low C+G contents, and we have demonstrated improved prediction accuracy for CG-poor genes by using this new training dataset compared to other methods. The combination of the two datasets enables GeneKey to get high and balanced prediction accuracy for both CG-rich and CG-poor genes.

Obviously, not all of the genes have similar structure features, because of the function restriction. Statistical analysis demonstrates that some structure features, such as splicing signals and codon usage, of CG-poor genes are quite different from that of CG-rich ones.

The results of this work imply that careful construction of training dataset is very important for uncovering underlying features and improving the performance of various prediction tasks.

Acknowledgements

This work was supported by the National Natural Science Foundation of China (Grants No. 90203011 and 30370354), the Science Foundation of Hubei Province of China (Grant No. 2002AC014), the Ministry of Education of China (Grant No. CG2003-GA002) and the Ministry of Science and Technology of China (Grant No. 2003AA231010).

References

[1] M. Q. Zhang. Computational Prediction of Eukaryotic Protein-coding Genes. *Nature Reviews Genetics*, 2002, 3: 698-709.

[2] C. Mathe, M. F. Sagot, T. Schiex, et al. Current Methods of Gene Prediction, Their Strengths and Weaknesses. *Nucleic Acids Research*, 2002, 30: 4103-4117.

[3] C. Burge, S. Karlin. Prediction of Complete Gene Structures in Human Genomic DNA. *Journal of Molecular Biology*, 1997, 268: 78-94.

[4] M. Q. Zhang. Identification of Protein Coding Regions in the Human Genome by Quadratic Discriminant Analysis. *Proc. Natl. Acad. Sci. USA*, 1997, 94: 565-568

[5] A. Krogh. Using Database Matches with HMMGene for Automated Gene Detection in Drosoplila. *Genome Research*, 2000, 10: 523-528.

[6] M. G. Reese, D. Kulp, H. Tammana, et al. Genie - Gene Finding in Drosophila Melanogaster. *Genome Research*, 2000, 10: 529-538.

[7] S. Batzoglou, L. Pachter, J. P. Mesirov, et al. Human and Mouse Gene Structure: Comparative Analysis and Application to Exon Prediction. *Genome Research*, 2000, 10: 950-958.

[8] P. M. Hooper, H. Zhang, D. S. Wishart. Prediction of Genetic Structure in Eukaryotic DNA Using Reference Point Logistic Regression and Sequence Alignment. *Bioinformatics*, 2000, 16: 425-438.

[9] R. F. Yeh, L. P. Lim, C. B. Burge. Computational Inference of Homologous Gene Structures in the Human Genome. *Genome Research*, 2001, 11: 803-816.

[10] I. Korf, P. Flicek, D. Duan, et al. Integrating Genomic Homology into Gene Structure Prediction. *Bioinformatics*, 2001, 17: S140-S148.

[11] L. Taher, O. Rinner, S. Garg, et al. AGenDA: homology-based gene prediction. *Bioinformatics*, 2003, 19: 1575-1577.

[12] Y. H. Zhou, L. Yang, H. Wang, et al. Prediction of Eukaryotic Gene Structures Based on multilevel Optimization. *Chinese Science Bulletin*, 2004, 49: 321-328.

[13] Y. H. Zhou, H. Jing, Y. E. Li, et al. Identification of True EST Alignments and Exon Regions of Gene Sequences. *Chinese Science Bulletin*, 2004, 49: 2463-2469.

[14] E. Lander, et al. Initial Sequencing and Analysis of the Human Genome. *Nature*, 2001, 409: 860-921.

[15] J. C. Venter, M. D. Adams, E. W. Myers, et al. The Sequence of the Human Genome. *Science*, 2001, 291: 1304-1351.

[16] M. Burset, R. Guigó. Evaluation of Gene Structure Prediction Programs. *Genomics*, 1996, 34: 353-367.

[17] S. Rogic, A. Mackworth, F. B. F. Ouellette. Evaluation of Gene-finding Programs on Mammalian Sequences. *Genome Research*, 2001, 11: 817-832.

[18] H. Wan, J. C. Wootton. A Global Compositional Complexity Measure for Biological

Sequences: AT-rich and GC-rich Genomes Encode Less Complex Proteins. *Comput. Chem.*, 2000, 24: 67- 88.

Yanhong Zhou is a professor in the School of Life Science and Technology, Huazhong University of Science and Technology (HUST), People’s Republic of China. He received his BS, MS, and PhD in mechanical engineering from HUST. From January 1999 to December 2001, he did two years of postdoctoral research at Harvard University, and one year of visiting research at University of Missouri-Columbia, USA. His current research interests are in the fields of bioinformatics and biomedical engineering.

Huili Zhang is a lector in the Center of Computing & Enperimenting at South-Central University for Nationalities. She received her BS in the School of Computer Science at South-Central University for Nationalities in 1996, and is now a graduate student in the School of Computer Science and Technology, Huazhong University of Science and Technology. Her current research interests are in the field of genome sequence analysis.

Lei Yang is a graduate student in the School of Computer Science and Technology, Huazhong University of Science and Technology (HUST), People’s Republic of China. He received his BS in physics from HUST in 2003. His current research interests are in the fields of genome sequence analysis.