July 2005, Vol.20, No.4, pp.446–453 J. Comput. Sci. & Technol.

Test Data Sets and Evaluation of Gene Prediction Programs on the Rice Genome

Heng Li[1,2*] (李 恒), Jin-Song Liu[1*] (刘劲松), Zhao Xu[1*] (徐 昭), Jiao Jin[1,3] (金 蛟), Lin Fang[1] (方 林), Lei Gao[1,2] (高 雷), Yu-Dong Li[1] (李余动), Zi-Xing Xing[1,3] (邢自兴), Shao-Gen Gao[1,4] (高绍根), Tao Liu[1] (刘 涛), Hai-Hong Li[1] (李海红), Yan Li[5] (李 雁), Li-Jun Fang[5] (方丽君), Hui-Min Xie[6] (谢惠民), Wei-Mou Zheng[1,2] (郑伟谋), and Bai-Lin Hao[2,5,7**] (郝柏林)

[1] *Beijing Genomics Institute (BGI), Academia Sinica, Beijing 101300, P.R. China*

[2] *Institute of Theoretical Physics, Academia Sinica, Beijing 100080, P.R. China*

[3] *Department of Mathematics, Beijing Normal University, Beijing 100875, P.R. China*

[4] *Institute of Systems Science, Academia Sinica, Beijing 100080, P.R. China*

[5] *Hangzhou Branch of BGI, Hangzhou 310008, P.R. China*

[6] *Department of Mathematics, Suzhou University, Suzhou 215006, P.R. China*

[7] *T-Life Research Center, Fudan University, Shanghai 200433, P.R. China*

E-mail: hao@itp.ac.cn

Received September 3, 2004; revised November 20, 2004.

Abstract With several rice genome projects approaching completion gene prediction/finding by computer algorithms has become an urgent task. Two test sets were constructed by mapping the newly published 28,469 full-length KOME rice cDNA to the RGP BAC clone sequences of *Oryza sativa* ssp. *japonica*: a single-gene set of 550 sequences and a multi-gene set of 62 sequences with 271 genes. These data sets were used to evaluate five *ab initio* gene prediction programs: RiceHMM, GlimmerR, GeneMark, FGENSH and BGF. The predictions were compared on nucleotide, exon and whole gene structure levels using commonly accepted measures and several new measures. The test results show a progress in performance in chronological order. At the same time complementarity of the programs hints on the possibility of further improvement and on the feasibility of reaching better performance by combining several gene-finders.

Keywords gene prediction, rice genome, test sets, accuracy measures, hidden Markov models, dynamic programming

1 Introduction

Rice is the most important staple food for more than half of the world population. It is critical for the sustained development of mankind to increase the production and to improve the quality of rice. At the same time the rice genome is the smallest and most compact among those of cereal crops. This makes rice a model organism for monocotyledonous plant as compared to dicotyledons such as *Arabidopsis thaliana*. Therefore, the sequencing of the rice genome is undebatably of primary interest. With several rice genome sequencing projects[1−3] approaching completion[4−7] finding genes in the rice genome has become an urgent task. However, in contrast to gene prediction in mammalian genomes relatively few programs have been devoted to plant genomes (see, e.g., a recent review on computational gene finding in plants[8]). The lack of commonly accepted test sets has also been felt as an impeding factor.

Computational gene finding started in the early 1980s, though workable gene prediction programs for eukaryotes appeared only in the 1990s, see the comprehensive review by Solovyev[9]. The first eukaryotic gene finders may be divided into *ab initio* and similarity based programs. The *ab initio* programs are based on gene structure models with parameters learned from training set and do not use similarity information from searching databases of known proteins, ESTs and cDNAs. One should admit that significant progress of *ab initio* gene prediction has not been observed in the last few years. Instead there have been much efforts in finding genes by comparative genomics (dual-genome predictors such as SGP-2, TWINSCAN, or SLAM as reviewed in [10]) or by combining the predictions of several gene-finders such as GeneComber[11] or Combiner[12].

However, comparative genomics gene prediction cannot be effectively used for the rice genome at present because the only other plant genome sequenced so far is that of *A. thaliana*. The latter is too distant from rice as evidenced by the great number of rice genes that do not have homologs in *A. thaliana* (so-called NH genes[1] was estimated to make 1/3 of the rice genes). Consequently, *ab initio* gene prediction programs based on a single genome necessarily remain important annotation tools for rice and the evaluation of these programs may be very instructive for their improvement. Furthermore, the feasibility of combining outputs of several gene predictors may be elucidated only after proper evaluation of the available programs. Therefore, we concentrate on

Regular Paper

* These three authors contributed equally.

** Corresponding author.

ab initio gene-finders for the rice genome and construct data sets for their evaluation.

2 Test Data Sets

For testing gene prediction in the human genomic sequences there exist a few frequently used data sets. The 1996 Burset-Guigó ALLSEG set contains 570 single-gene multi-exon sequences[13]. The 2000 Guigó set h178[14] contains 178 single-gene sequences and a semi-artificial multi-gene set SAG42 was constructed by using the h178 genes and inserting random "intergenic" segments. The 2001 HMR195 set[15] emphasized on avoiding overlaps with the training data. These data sets have been in use until recent time. We note that there has not been a test set of natural and verified multi-gene sequences.

As for plant genomes Kleffe *et al.*[16] constructed two sets GBEzm and GBEat. The former contains 46 maize genes with 250 exons and 204 introns and the latter contains 131 *A. thaliana* genes with 709 exons and 578 introns. These data were produced even before the *A. thaliana* genome was completely sequenced. The genes in *A. thaliana* were predicted[17] by combining several programs (GeneMark, XGrail, GeneFinder, GENSCAN and Netplantgene) and tested by an unpublished collection of sequences with 100 experimentally verified genes. The training data set of the rice gene-finder GlimmerR (see below) was described in [18] but it was tested on the training data and on a single GenBank sequence not included in the training data.

Therefore, the construction of high-quality test sets for the rice genome remains an actual and urgent task. We address this task in what follows. We first characterize the two test sets OsSNG550 and OsMTG62 and then describe how they are obtained and how well they represent the "bulk" data. These data sets are available from the BGI-RIS web site[7].

2.1 Single-Gene Data Set

The single-gene set OsSNG550 contains 550 genomic sequences of length 1,093 to 10,708 bases. Each sequence contains a cDNA-verified gene. There are 431 multi-exon genes and 119 single-exon genes. The total number of exons are 2,534. All the 1,984 introns have the canonical minimal splicing signal GT—AG to fit the present-day gene prediction programs that ignore non-canonical splicing.

2.2 Multi-Gene Data Set

The multi-gene set OsMTG62 contains 62 sequences of length 10,813 to 42,698 bases with a total of 271 genes. Each sequence contains 4 to 8 genes. There are 53 single-exon genes among the 271.

We note that this is the first set of natural, i.e., not semi-artificial, multi-gene test sequences ever constructed for eukaryotic genomes.

2.3 Construction of the Data Sets

The rules we followed in constructing a test set are listed below:

1) it consists of experimentally, i.e., cDNA confirmed genes;

2) the characteristics of the set is representative of the bulk data on average;

3) the collection of sequences has minimal overlap with possible training data of the programs to be tested.

The publication of the KOME[19] full-length rice cDNA provides an ideal opportunity to construct test sets to satisfy the above requirements as stringent as possible. Though said to be full-length, the KOME cDNAs may still contain erroneous or unexpected factors (see, e.g., recent comments[20]) which are not considered in the common gene-finding software. The current gene prediction programs are designed to detect "standard" protein-coding genes. Before all the KOME cDNAs are fully annotated and understood, we must filter the KOME data to get a reliable subset to commence with.

We start from the 28,469 cDNAs and proceed as follows.

1) Seek the best Open Reading Frame (ORF) by dynamic programming for each cDNA, assuming that there is always a complete gene structure including the start and stop codons, regardless of the fact that a cDNA may just be a non-coding gene or various flaws may distort its structure as a coding gene. A cDNA will be discarded if the longest ORF is less than 300bp. A total of 21,545 cDNAs remain after this step.

2) Map these cDNAs with ORF determined to the BAC sequences downloaded from GenBank in February 2003. Those BACs including long runs of Ns were split at these gaps in order to insure that possible introns would not contain Ns. We require more than 95% match of the full cDNA length. A cDNA will be discarded if part of its CDS drops off the aligned region or if there are indels not located in prospective introns. All short "introns" less than 60 bases were considered as indels and discarded. In other words minimal intron length in what left is 61 bases. Then we check for redundancy by using BLAT[21]. If two cDNAs overlap more than 100 consecutive bases the shorter one is discarded. A total of 10,697 sequences remain after this step.

3) Since 5% mismatch was allowed in the above step, there were sequences which contain in-frame stop codons, incorrect start or stop codons, non-canonical (i.e., not GT—AG) splicing sites, etc. These sequences were dropped. In addition, a few sequences with possible alternative splicing were also discarded. In order to guarantee that the test data do not overlap with possible training data of the programs to be tested we aligned the remaining cDNAs with all rice cDNAs in

GenBank Rel. 132 (October 15, 2002) using BLAT again. Any cDNA with more than 100-base overlap with known cDNA was dropped. A total of 7,455 single-gene sequences were kept after this step.

4) Then we proceeded to insure that these cDNAs do correspond to some known proteins to a certain extent. First the translated cDNAs were aligned against the proteins of *A. thaliana* by BLAST[22] using a cutoff E-value less than 10^{-7}. We found 5,548 cDNAs with homolog to the *A. thaliana* proteins and 1,907 without homolog to the latter. In terms of the abbreviation used in [1] the 7,455 cDNAs consist of 5,548 WH genes and 1,907 NH genes. From the 5,548 WH genes 500 were randomly sampled to make a major part of the OsSNG550 data set.

5) Among the 1,907 NH genes 885 were further aligned with other plant proteins, 540 with proteins of other non-plant organisms and 482 had no alignment. Then 50 genes were randomly sampled from the 1,425 (= 885 + 540) sequences and added to the OsSNG550 set.

The construction of the set of multi-gene sequences had to be started with the 10,697 sequences left after the second step above. Our main concern was that there should not be actual genes in the seemingly "intergenic" regions as the collection of KOME cDNAs covers only a limited portion of rice genes. We took all "intergenic" regions from the multi-gene sequences. These were first aligned with the 123,443 public ESTs of the NCBI dbEST (October 2002) and all known rice cDNAs (GenBank Rel. 132). When matched bases are greater than 90% of the sequence length it is disqualified. Then the remaining sequences were aligned with all rice proteins and *A. thaliana* proteins in GenBank Rel. 132 using BLAST with an $E = 10^{-7}$ cutoff. Only 1,318 multi-gene sequences were left of which 62 sequences contain more than 4 genes and they comprise our OsMTG62 set. Due to the limited number of sequences we have to be content with the fact that some non-canonical splicing sites remain. In fact, there are 20 non-canonical sites in 15 genes. These are slightly higher than the ratio of non-canonical splicing in the bulk data (about 1%).

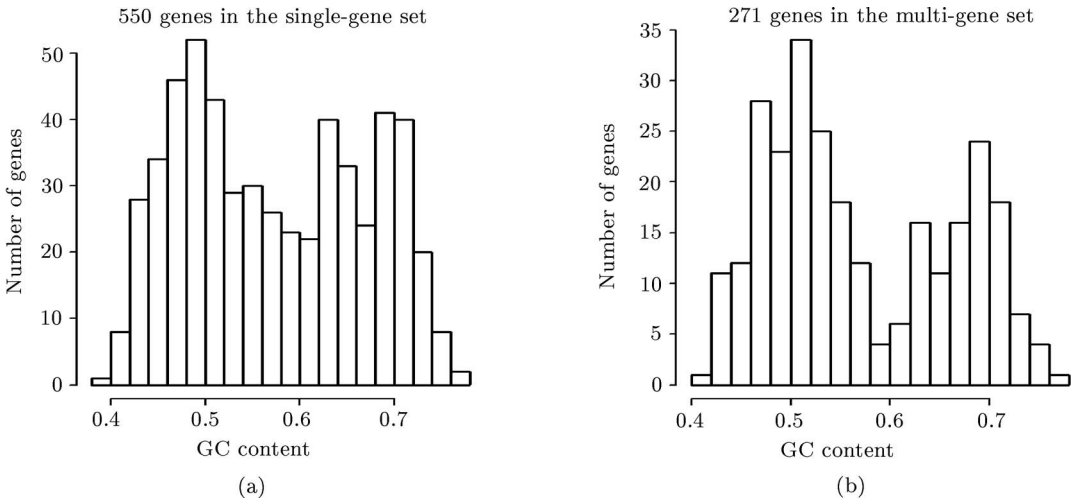

Fig.1. GC content of the test sets. (a) OsSNG550 set. (b) OsMTG62 set.

Table 1. Comparison of Exon and Intron Characteristics in the Bulk Data and Test Sets

Data sets	Bulk data		Test sets	
	KOME WH	KOME NH	OsSNG550	OsMTG62
Seq number	5,548	1,425	550	62
Gene Number	5,548	1,425	550	271
Intron Number	22,293	3,984	1,984	1,102
max length	19,575	12,171	15,100	8,143
mean length	395	387	426	380
Exon Number	27,841	5,409	2,534	1,373
max length	4,284	4,032	2,742	3,462
mean length	247	273	255	243
Init Exon Number	4,374	942	431	218
max length	3,907	2,286	2,742	2,412
mean length	319	316	327	299
Intr Exon Number	17,919	3,042	1,553	884
max length	2,646	2,290	2,646	1,619
mean length	150	159	151	154
Term Exon Number	4,374	942	431	218
max length	3,730	2,174	2,390	1,649
mean length	321	295	322	298
Sngl Exon Number	1,174	483	119	53
max length	4,284	4,032	2,685	3,462
mean length	1,170	862	1,114	1,265

Keeping a few non-canonical splicing sites makes the data more realistic as true biological sequences are not that ideal.

2.4 Average Characteristics of the Test Data Sets

In order to show that the test sets do represent the "bulk" data on average we consider the GC content and intron/exon size of the genes in the test data.

It was first observed in the rice draft genome that the exon GC content distribution showed a two-peak feature, see Fig.3 in [1]. This has been further verified on the collection of more predicted genes and the KOME cDNAs. Fig.1 shows the distribution of genes by their GC content for the 550 and 271 genes in the OsSNG550 and OsMTG62 data sets respectively. Both curves do exhibit the two-peak feature.

The comparison of exon and intron characteristics in the test sets with that of the bulk data is given in Table 1. Here the two groups of bulk data refer to the 5,548 WH genes and the 1,425 NH genes that have been selected from the 28,469 KOME[19] cDNAs as described above. The test data do represent the bulk data on average.

3 Programs Tested

We compare five gene prediction programs: RiceHMM, GlimmerR, GeneMark.hmm, FGENESH, and BGF. All these programs look for multiple genes on both DNA strands. They are summarized in Table 2 and a short description of each program follows.

Table 2. Programs Used in This Work

Program	Version	Trained on	Last update
RiceHMM		Rice	Aug. 2002
GlimmerR	1.0	Rice	2001
GeneMark	2.2a	Rice	May 2002
FGENESH	2.0	Monocots	2002
BGF	1.01	Rice	Aug. 2003

RiceHMM is "based on probabilistic model using a catalog of rice ESTs" (Sakata *et al.*, 1999)[23]. It was trained on rice data and the algorithm and parameters of RiceHMM was last updated in August 2002. A web service is available at http://rgp.dna.affrc.go.jp/RiceHMM/index.html

GlimmerR[18] is a specific version of eukaryotic gene-finder GlimmerM[24] trained specially for rice. GlimmerM was originally developed for *Plasmodium falciparum* by modifying the microbial gene identification program Glimmer 2.0[25] which was based on Interpolated Markov models. GlimmerR may be accessed via the GlimmerM webpage at http://www.tigr.org/tdb/glimmerm/glmr_form.html

GeneMark.hmm (Borodovsky and Lukashin, unpublished) is an HMM-based program as its name hints on. It runs in parallel with the eukaryotic version of GeneMark[26] which was first developed for gene-finding in prokaryotic genomes. GeneMark.hmm was updated for the rice genome on May 10, 2002. The web page of GeneMark.hmm is http://opal.biology.gatech.edu/GeneMark/eukhmm.cgi

FGENESH[27] is an HMM-based *ab initio* gene structure prediction program. It was described in [9] and references therein. The version 2.0 we used was trained on monocotyledons (corn, rice, wheat, and barley), but the training details are not available. The Syngenta group[2] used it to produce 87% of all high-evidence predicted genes and the BGI group[1] estimated it as "by far the most accurate of 5 programs tested" on the rice draft genome when BGF was still under development. The FGENESH home page is http://www.softberry.com/berry.phtml

BGF (Beijing Gene Finder) is based on hidden semi-Markov model and dynamic programming. Though having an overall structure similar to that of GENSCAN many distinctive features have been added in its implementation. For example, its signal search and enhancement strategy has been described in [28–30]. BGF is written in C++ using the Standard Template Library (STL). Evaluation version of BGF has been made public since the publication of the BGI-RIS[7] — the Rise Information System RIS^e of BGI. It was trained on public data available by the end of 2002, producing parameters for two GC isochores. BGF has also been trained for the silkworm *Bombyx mori* genome. Genomic sequences may be submitted to BGF for gene-finding at http://bgf.genomics.org.cn/

We did not include GENSCAN[31,32] in the comparison as there was only a version trained on early maize data. However, GENSCAN was perhaps one of the first and best *ab initio* hidden semi-Markov model and dynamic programming based gene-finders for the human genome and its architecture has been instructive for many subsequent gene prediction programs, including RiceHMM, FGENESH and BGF. It is worth mentioning that when tested on rice genomic sequences GENSCAN's overall performance on multi-gene sequences is quite impressive, see Table 10 below. GENSCAN's homepage is http://genes.mit.edu/GENSCAN.html

4 Prediction Accuracy Measures and Test Results

We evaluate the accuracy of gene predictions at four levels: the nucleotide level, the exon level, the whole-gene structure level, and the multi-gene sequence level.

4.1 Nucleotide Level

At the nucleotide level there are some commonly accepted measures[13,16,33] which are defined without ambiguity. We follow these definitions and list them for reference. A prediction is compared with the actual situation in the test set base by base. Nucleotides may be predicted as coding or non-coding. The total number of

Table 3. Test Results at the Nucleotide Level

Program	Single-gene set OsSNG550			Multi-gene set OsMTG62		
	S_n	S_p	CC	S_n	S_p	CC
RiceHMM	0.70	0.81	0.66	0.69	0.68	0.61
GlimmerR	0.69	0.83	0.66	0.61	0.64	0.54
GeneMark	0.91	0.89	0.85	0.91	0.79	0.80
FGENESH	0.96	0.93	0.92	0.96	0.82	0.86
BGF	0.97	0.94	0.93	0.96	0.83	0.86

Table 4. Predicted Number of Exons in Each Class in the Single-Gene Data Set OsSNG550. AE = Actual Exon number. The ratios given in the table are the TE/PE values.

Class	Initial	Internal	Terminal	Single	Total
AE	431	1,553	431	119	2,534
RiceHMM	123/544	321/964	124/600	17/105	585/2,213
GlimmerR	173/450	569/751	143/450	56/415	941/2,066
GeneMark	222/606	1,237/1,829	185/336	36/72	1,680/2,843
FGENESH	288/408	1,363/1,674	322/442	72/107	2,045/2,631
BGF	302/411	1,372/1,679	315/389	64/87	2,053/2,566

of coding nucleotides is denoted by PP (Predicted Positive), that of non-coding ones contributes to PN (Predicted Negative). The corresponding numbers in the test set are AP (Actual or Annotated Positive) and AN (Actual or Annotated Negative). If a predicted base actually falls in a coding segment it is counted as TP (True Positive), otherwise it is FP (False Positive). Similarly we have TN (True Negative) or FN (False Negative). We have clearly

$$AP = TP + FN, \quad AN = TN + FP;$$
$$PP = TP + FP, \quad PN = TN + FN.$$

Then sensitivity S_n and specificity S_p of the predictions are defined as

$$S_n = \frac{TP}{AP} = \frac{TP}{TP + FN}, \quad S_p = \frac{TP}{PP} = \frac{TP}{TP + FP}.$$

In order to measure the global performance of a gene-finder at the nucleotide level one can use either the *correlation coefficient CC*

$$CC = \frac{(TP)(TN) - (FP)(FN)}{\sqrt{(PP)(PN)(AP)(AN)}} \tag{1}$$

or the *approximate correlation AC*

$$AC = \frac{1}{2}\Big(\frac{TP}{TP + FN} + \frac{TP}{TP + FP} + \frac{TN}{TN + FP} + \frac{TN}{TN + FN}\Big) - 1.$$

Both CC and AC yield $+1$ when the predictions are correct, i.e., $FP = FN = 0$, and lead to -1 when the predictions are entirely wrong, i.e., $TP = TN = 0$. Since it was observed that AC and CC are quite close in most cases and CC has a probabilistic interpretation[13] we use CC in this paper.

Test results on the nucleotide level are shown in Table 3 for the single-gene set OsSNG550 and multi-gene set OsMTG62. For all the programs the performance on the multi-gene set is slightly lower than that on the single-gene set.

4.2 Exon Level

The prediction accuracy at exon level is more important from a practitioner's point of view, e.g., for primer or probe design. We know the number of Actual exons (AE) from the annotated test set. The number of predicted exons (PE) comes directly from the program. However, it is more subtle to define a true exon (TE), since unlike the nucleotide case an exon prediction may often be partially correct. We take a stringent attitude in defining TE, namely, we only count an exon as TE when it has both beginning and ending positions and, naturally, the length, coinciding with that of the actual one. Having AE, PE, and TE at hand we define sensitivity ES_n and specificity ES_p at the exon level:

$$ES_n = \frac{TE}{AE}, \qquad ES_p = \frac{TE}{PE}.$$

Table 5. Sensitivity and Specificity of Predictions for Various Classes of Exons in the Single-Gene Data Set OsSNG550

Program	Initial		Internal		Terminal		Single		Total	
	ES_n	ES_p	ES_n	ES_p	ES_n	ES_p	ES_n	ES_p	ES_n	ES_p
RiceHMM	0.28	0.15	0.21	0.23	0.26	0.15	0.13	0.13	0.23	0.26
GlimmerR	0.24	0.22	0.33	0.68	0.24	0.22	0.45	0.08	0.37	0.46
GeneMark	0.61	0.47	0.72	0.63	0.60	0.48	0.66	0.51	0.66	0.59
FGENESH	0.69	0.61	0.84	0.71	0.72	0.72	0.70	0.54	0.81	0.78
BGF	0.70	0.63	0.85	0.76	0.76	0.76	0.75	0.49	0.81	0.80

Table 6. Predicted Number of Exons in Each Class in the Multi-Gene Data Set OsMTG62. AE = Actual Exon Number. The ratios given in the table are the TE/PE values.

Class	Initial	Internal	Terminal	Single	Total
AE	218	884	218	53	1,373
RiceHMM	61/395	186/804	57/392	7/52	311/1,643
GlimmerR	52/236	293/431	52/236	24/299	421/1,202
GeneMark	134/285	637/1,016	130/273	35/689	936/2,263
FGENESH	150/244	741/1,051	157/244	37/69	1,085/1,608
BGF	152/240	754/998	166/239	40/82	1,112/1,559

Table 7. Sensitivity and Specificity of Predictions for Various Classes of Exons in the Multi-Gene Data Set OsMTG62

Program	Initial		Internal		Terminal		Single		Total	
	ES_n	ES_p	ES_n	ES_p	ES_n	ES_p	ES_n	ES_p	ES_n	ES_p
RiceHMM	0.29	0.23	0.21	0.33	0.29	0.21	0.14	0.16	0.23	0.19
GlimmerR	0.40	0.38	0.37	0.76	0.33	0.32	0.47	0.13	0.31	0.35
GeneMark	0.62	0.37	0.80	0.68	0.43	0.55	0.30	0.50	0.68	0.57
FGENESH	0.69	0.71	0.88	0.81	0.75	0.76	0.61	0.67	0.79	0.67
BGF	0.70	0.73	0.88	0.82	0.73	0.81	0.54	0.74	0.81	0.71

For the time being exons are subdivided into four classes: initial, internal, and terminal exons in multi-exon genes plus exons in single-exon genes. Due to lacking of training data and statistical analysis introns and exons in the 5′-UTR and 3′-UTR are not treated in the programs tested, not to mention the proposal to divide exons into 12 classes[34]. Since the performance of gene-finders differs for various classes we list separately the *AE*, *PE*, and *TE* in Tables 4 and 6 for the OsSNG550 and OsMTG62 sets, respectively.

Table 8. Wrong and Missing Exons in the Two Test Sets

Program	OsSNG550 set		OsMTG62 set	
	WE	*ME*	*WE*	*ME*
RiceHMM	0.25	0.37	0.46	0.38
GlimmerR	0.20	0.35	0.37	0.43
GeneMark	0.17	0.09	0.26	0.12
FGENESH	0.09	0.05	0.21	0.07
BGF	0.07	0.05	0.18	0.06

Using the *AE*, *TE* and *PE* values given in these tables one can easily calculate the exon sensitivity ES_n and exon specificity ES_p for each class of exons. Given in Tables 5 and 7 are these measures for exons in the two data sets OsSNG550 and OsMTG62, respectively.

We note that the rice genome has much more single-exon genes (from 20% in low-GC genes to 40% in high-GC genes) than the human genome.

Not trying to define a mutually exclusive set of various partially correct exons, we emphasize only on the number of *Wrong Exons* and the number of *Missing Exons*. An actual exon is counted as missing if it does not have a single base predicted. A predicted exon is counted as wrong if no single predicted base is present in the actual exons. These are again stringent, all or none, measures. To be precise, we define

$$ME = \frac{\text{No. of missing exons}}{\text{No. of actual exons}},\ WE = \frac{\text{No. of wrong exons}}{\text{No. of predicted exons}}.$$

The scores *WE* and *ME* are given in Table 8.

4.3 Whole Gene Structure Level

At the whole gene structure level there are many more diverse cases of partially correct predictions. For example, actual genes may be split or joined in the predictions. Instead of attempting to define a mutual exclusive set of partially correct predictions we again take a very stringent attitude to evaluate the performance of gene-finders at the whole gene structure level. A predicted gene is said to be a *Right Gene* (*RG*) only if it coincides with the annotation 100% with all the start, stop, and splicing sites correctly determined as compared to the cDNA mapping to genomic sequences. Likewise a gene is considered as a *Missing Gene* (*MG*) if no single base has been predicted. The results are summarized in Tables 9 and 10 for the single-gene set OsSNG550 and multi-gene set OsMTG62, respectively.

Table 9. Number of Correctly Predicted (*RG*), Partially Predicted (*PG*) and Missing (*MG*) Genes Among the 550 Actual Genes in the Single-Gene Set OsSNG550

Program	*RG*	*PG*	*MG*
BGF	237 (37)	308	5
FGENESH	231 (28)	315	4
GeneMark	116 (16)	418	16
GlimmerR	86 (21)	453	11
RiceHMM	44 (3)	492	14

The numbers given in parentheses in the first columns in Tables 9 and 10 show the numbers of right genes predicted only by the specified gene-finder only, but not by any other programs. This shows the complementarity of these programs and indicates on the feasibility of combining predictions from different gene-finders to improve the overall performance. If one count the total number of right genes predicted by at least one program, it is 332 and 163 for OsSNG550 and OsMTG62, respectively. It is interesting to note that $332/550 \approx 163/271 \approx 0.60$.

4.4 Test on Multi-Gene Sequence Level

Having a multi-gene test set at hand we are in a position to evaluate the gene prediction on the whole sequence level. In accordance with our stringent requirement among the 62 multi-gene sequences we count the number of sequences whose all genes were predicted entirely correct. Although there is no guarantee that the regions annotated as "intergenic" are free of genes we assume that they are all true intergenic regions. Among the 62 sequences only 5 were correctly predicted by one or more programs. The results are given in Table 11 where 1 stands for correct prediction.

An example of gene prediction at the multi-gene sequence level is given in Fig.2. This figure was produced by using the gff2ps software[35]. Note that the arrows indicate the direction of transcription.

In principle, one can go on to test the gene-finders on un-annotated genomic sequences. Since there is no objective reference for the accuracy of predictions on these sequences we have to rely on mutual consistency of the programs. To prepare for such test in the future we study the mutual consistency of the six programs on the single-gene set OsSNG550.

Table 10. Number of Correctly Predicted (*RG*), Partially Predicted (*PG*) and Missing (*MG*) Genes Among the 271 Actual Genes in the Multi-Gene Set OsMTG62

Program	*RG*	*PG*	*MG*
BGF	124 (14)	145	2
FGENESH	114 (8)	155	2
GeneMark	89 (9)	180	2
GlimmerR	32 (4)	219	20
GENSCAN	31 (5)	221	19
RiceHMM	24 (0)	232	15

Table 11. Correct Predicted Sequences in the OsMTG62 Set

Sequence	RiceHMM	GlimmerR	GeneMark	FGENESH	BGF
MT000018	0	0	1	0	1
MT000064	0	0	0	0	1
MT000636	0	0	1	1	1
MT000823	0	0	0	1	1
MT000943	0	0	0	1	1

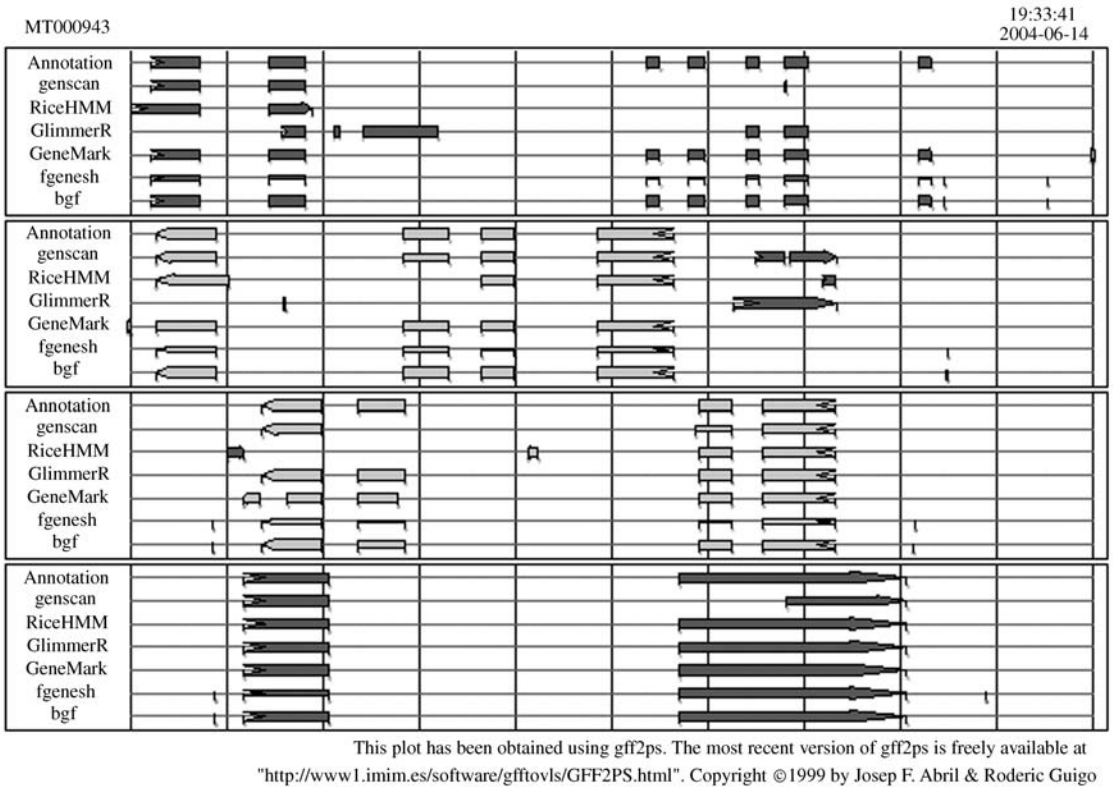

Fig.2. Comparison of gene prediction by the six gene-finders on the test genomic sequence MT000943 in the OsMTG62 data set that contains 4 genes. The annotation line shows the structure of the actual genes.

We calculate the correlation coefficient CC as defined in (1) at the nucleotide level for each pair of programs treating one of them as the correct reference. The CCs with the annotation are also included. The results are shown in Table 12. This is essentially a 7×7 symmetric normalized correlation matrix (we have included GENSCAN for completeness). Then a normalized distance D between two programs is defined as $D = 1 - CC$.

Table 12. Pairwise Correlation Coefficient (*CC*)

	Annotation	BGF	FGENESH	GeneMark	GENSCAN	GlimmerR	RiceHMM
Annotation	1.00	0.93	0.92	0.85	0.60	0.66	0.66
BGF	0.93	1.00	0.95	0.87	0.62	0.65	0.68
FGENESH	0.92	0.95	1.00	0.86	0.61	0.64	0.67
GeneMark	0.85	0.87	0.86	1.00	0.63	0.61	0.65
GENSCAN	0.60	0.62	0.61	0.63	1.00	0.49	0.59
GlimmerR	0.66	0.65	0.64	0.61	0.49	1.00	0.54
RiceHMM	0.66	0.68	0.67	0.65	0.59	0.54	1.00

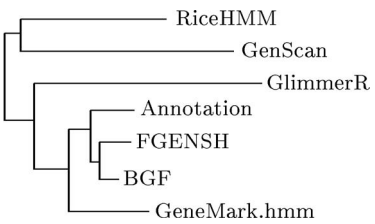

Fig.3. Relation among the 6 gene-finders and the actual situation, i.e., the annotation, as defined by their correlation distance (see text).

Using the distance matrix thus obtained we construct a "phylogenetic" tree of the programs and the annotation as shown in Fig.3. We emphasize that the closeness of two programs on the tree does not necessarily mean correctness of their predictions as common false predictions would also increase the CC and reduce the distance between programs. Nevertheless, Fig.3 agrees qualitatively with the numerical evaluation carried out in the previous sections.

5 Discussion

Summarizing the test results in the last section we see that there has been a steady progress of *ab initio* gene-predicting programs in chronological order. Moreover, the five gene-finders may be grouped in two sets: GeneMark.hmm, FGENESH and BGF have significantly better performance than RiceHMM and GlimmerR. Roughly speaking, on the rice genome the accuracy of gene prediction reaches 90% at nucleotide level, 80% at exon level, and nearly 60% at the whole-gene level as compared to 80%, 45%, and 20%, respectively, for the human genome[34]. We note also that rice gene expression experiments at the Beijing Genomics Institute have revealed genes which have not been supported by either EST or cDNA data, but do have been predicted by gene-finding programs. The seemingly better performance of gene-predicting software on rice genome is mainly caused by the smaller size and compactness of the rice genome.

However, all these *ab initio* programs have their drawbacks and limitations. The rich regulatory signals in the flanking regions of genes are not fully taken into account in locating the genes. They are not designed to deal with pseudogenes and transposons, non-canonical and alternative splicing sites, problems caused by frameshifts, etc. Their findings are necessarily biased towards the type of genes in the training sets. In addition, the complementarity of these programs also tells about the importance of score balancing — both short signal scores and long segmental scores — as the subtle differences in the number of the 100% correctly predicted genes may come from slight different ways of score-balancing. Obviously, there is space for improvement of *ab initio* gene-finding programs.

Acknowledgements The authors thank the BGI Rice Genome Sequencing Team, especially, Jun Wang, Pei-Xiang Ni, Xi-Miao He, and Chen Ye, for their constant support and interaction.

References

[1] Yu J, Hu S-N *et al.* A draft sequence of the rice genome (*Oryza sativa* L. ssp. *indica*). *Science*, 2002, 296: 79–92.

[2] Goff S A, Ricke D *et al.* A draft sequence of the rice genome (*Oryza sativa* L. ssp. *japonica*). *Science*, 2002, 296: 92–100.

[3] The international rice genome sequencing project. http://rgp.dna.affrc.go.jp/IRSGP/

[4] Sasaki T, Matsumoto T, Yamamoto K *et al.* The genome sequence and structure of rice chromosome 1. *Nature*, 2002, 420: 312–316.

[5] Feng Q, Zhang Y J, Wang S Y *et al.* Sequence and analysis of rice chromosome 4. *Nature*, 2002, 420: 316–320.

[6] The rice chromosome 10 sequencing consortium. In-depth view of structure, activity and evolution of rice chromosome 10. *Science*, 2003, 300: 1566–1569.

[7] Zhao W-M, Wang J, He X-M *et al.* BGI-RIS: An integrated information resource and comparative analysis workbench for rice genomics. *Nucl. Acids Res.*, 2004, 32: D377–D382.

[8] Pertea M, Salzberg S L. Computational gene finding in plants. *Plant Mol. Biol.*, 2002, 48(1): 39–48.

[9] Solovyev V V. Finding Genes by Computer: Probabilistic and Discriminative Approaches. Current Topics in Computational Molecular Biology, Jiang T, Xu Y, Zhang M Q (eds.), Tsinghua University Press and MIT Press, 2002, pp.201–248.

[10] Brent M R, Guigó R. Recent advances in gene structure prediction. *Curr. Opin. Struct. Biol.*, 2004, 14: 264–272.

[11] Shah S P, McVicker G P, Mackworth A K *et al.* GeneComber: Combining outputs of gene prediction programs. *Bioinformatics*, 2003, 9(10): 1296–1297.

[12] Allen J E, Pertea M, Salzberg S L. Computational gene prediction using multiple sources of evidence. *Genome Res.*, 2004, 14(1): 142–148.

[13] Burset M, Guigó R. Evaluation of gene structure prediction programs. *Genomics*, 1996, 34: 353–367.

[14] Guigó R, Agarwal P, Abril J F *et al.* An assessment of gene prediction accuracy in large DNA sequences. *Genome Res.*, 2000, 10(10): 1631–1642.

[15] Rogic S, Mackworth A K, Ouellette B F. Evaluation of gene-finding programs on mammalian sequences. *Genome Res.*, 2001, 11(5): 817–832.

[16] Kleffe J, Hermann K, Vahrson W *et al.* Logitlinear models for the prediction of splics sites in plant pre-mRNA sequences. *Nucl. Acids Res.*, 1996, 24(23): 4709–4718.

[17] The European Union Arabidopsis Genome Sequencing Consortium and the Cold Spring Harbor Washington University in St Louis and PE Biosystem Arabidopsis Sequencing Consortium. Sequence and analysis of chromosome 4 of the plant *Arabidopsis thaliana*. *Nature*, 1999, 402: 769–777.

[18] Yuan Q, Quackenbush J, Sultana R *et al.* Rice bioinformatics. Analysis of rice sequence data and leveraging the data to other plant species. *Plant Physiol.*, 2001, 125: 1166–1174.

[19] The Rice Full-Length cDNA Consortium. Collection, mapping and annotation of over 28,000 cDNA clones from *japonica* rice. *Science*, 2003, 301: 376–379.

[20] Jabbari K, Cruveiller S, Clay O *et al.* The new genes of rice: A closer look. *Trends in Plant Sci.*, 2004, 9(6): 281–285.

[21] Kent W J. BLAT: The BLAST-like alignment tool. *Genome Res.*, 2002, 12(4): 656–664.

[22] Altschul S F, Madden T L, Schaffer A A *et al.* Gapped BLAST and PSI-BLAST: A new generation of protein database search programs. *Nucl. Acid Res.*, 1997, 25(17): 3389–3402.

[23] Sakata K, Nagasaki H *et al.* A computer program for prediction of gene domain on rice genome sequence. In *The 2nd Georgia Tech Int. Conf. Bioinformatics Abstracts,* 1999, 78.

[24] Salzberg S L, Pertea M, Delcher A L *et al.* Interpolated Markov models for eukaryotic gene finding. *Genomics*, 1999, 59(1): 24–31.

[25] Delcher A L, Harmon D, Kasif S *et al.* Improved microbial gene identification with Glimmer. *Nucl. Acids Res.*, 1999, 27(23): 4636–4641.

[26] Borodovsky M, McIninch J. GENMARK: Parallel gene recognition for both DNA strands. *Computer Chem.*, 1993, 17(2): 123–133.

[27] Salamov A A *et al.* Ab initio gene finding in *Drosophila* genomic DNA. *Genome Res.*, 2000, 10(4): 516–522.

[28] Zheng W-M. Finding signals for plant promoters. *Genomics, Proteomics & Bioinformatics*, 2003, 1(1): 68–73.

[29] Zheng W-M. Genomic signal enhancement by clustering. *Commun. Theor. Phys.*, 2003, 39(5): 631–634.

[30] Zheng W-M. Genomic signal search by dynamic programming. *Commun. Theor. Phys.*, 2003, 39(6): 761–764.

[31] Burge C, Karlin S. Prediction of complete gene structures in human genomic DNA. *J. Mol. Evol.*, 1997, 268(1): 78–94.

[32] Burge C. Identification of genes in human genomic DNA [Dissertation]. Stanford University, 1997.

[33] Snyder E E, Stormo G D. Identification of coding regions in genomic DNA sequences: An application of dynamic programming and neural networks. *Nucl. Acids Res.*, 1993, 21: 607–613.

[34] Zhang M Q. Computational prediction of eukaryotic protein-coding genes. *Nature Reviews Genetics*, 2002, 3: 698–709.

[35] Abril J F, Guigó R. gff2ps: Visualizing genomic annotations. *Bioinformatics*, 2000, 16(8): 743–744.