Available online at www.sciencedirect.com
ScienceDirect

Agricultural Sciences in China
2008, 7(4): 387-394

April 2008

The Prediction of Rice Gene by Fgenesh

ZHANG Sheng-li[1, 2], LI Dong-fang[3], ZHANG Gai-sheng[1], WANG Jun-wei[1] and NIU Na[1]

[1] *Shaanxi Key Laboratory of Crop Heterosis, Northwest A & F University, Yangling 712100, P.R.China*
[2] *Key Discipline Open Laboratory on Crop Molecular Breeding, Henan Institute of Higher Learning/Henan Institute of Science and Technology, Xinxiang 453003, P.R.China*
[3] *School of Resources and Environmental Sciences, Henan Institute of Science and Technology, Xinxiang 453003, P.R.China*

Abstract

This study has been carried out to give some scientific reasons for genome annotation, shorten the annotating time, and improve the results of gene prediction. Taking the sequence of the 6th chromosome, which has more length sequences than others, of *Oryza sativa* L. ssp. *japonica cv.* Nipponbare as analysis data in this research, the gene prediction of monocots module, rice, has been done by using Fgenesh ver. 2.0, and the predicting results have been explored particularly by bioinformatics methods. Results showed that the number of predicted genes for this chromosome was very close to the number of TIGR annotated genes. The majority of the predicted genes were multi-exon genes which had a percentage of 77.52. Length range was very big in the predicted genes. According to the significant match number, multi-exon genes can be predicted more veracity than single exon genes and the support can be reached up to 100% by TIGR annotation and up to 78% by cDNA. From the angle of predicted exons location of multi-exon genes, the internal exons and last exons had a high support of cDNA. The length of internal exons was relatively short in high (>95% length, >78% similarity) cDNA and/or TIGR annotation support multi-exon genes, but the first exons and last exons were on the reverse. The majority of single exon genes which had more than 95% in length, and 78% in similarity support by cDNA and/or TIGR annotation was relatively short in length. From the angle of exon number, the majority of the multi-exon genes of high (>95% length, >78% similarity) cDNA and/or TIGR annotation support had no more than 5 exon number. It was concluded that the rice gene prediction by Fgenesh was very good but needed modification manually to some extent according to cDNA support after aligning the predicting sequence of genes with cDNA database of rice.

Key words: rice, gene prediction, cDNA, annotation, exon

INTRODUCTION

More than 600 species have been sequenced (Fig.1) and as much as 1 676 species are being sequenced according to the statistics of GOLD (http://igweb.integratedgenomics.com/GOLD/). By Jan. 2007, there were 716 eukaryote GSPs (genome sequencing projects) (31.18% of all GSPs) and 130 plant GSPs (18.16% of all eukaryotic GSPs) in the world. Astronomical sequence data have been provided along with a great deal of GSPs completion. How to annotate for these different biological sequences is the first task. Gene sequences, repeat sequences, and gene-spacing sequences (included intron sequences) constitute genomic sequences. Although gene sequences may take only a very small proportion in high eukaryote genomes, as the key factor controlling biological genetic character,

This paper is translated from its Chinese version in *Scientia Agricultura Sinica.*
ZHANG Sheng-li, Associate Professor, Ph D, E-mail: slzhang2008@yahoo.com.cn; Correspondence ZHANG Gai-sheng, Tel: +86-29-87092292, E-mail: zhanggsh@public.xa.sn.cn

accurate gene annotation from genomic sequences looks very important. Computational gene annotation is the first step after sequencing. So a lot of softwares can be used for gene annotation, but the Fgenesh (Salamov and Solovyev 2000), Genscan (Burge and Karlin 1997), Glimmer (Majoros *et al.* 2003), and GeneMark (Lomsadze *et al.* 2005), etc. are more accurate than others for the high eukaryotes.

Several researchers have evaluated the above software for rice gene annotation and deemed Fgenesh is the best one (Yu *et al.* 2002; Pertea and Salzberg 2002; Feng *et al.* 2002; Goff *et al.* 2002). But, they used only short sequences (400 kb-2.89 Mb) or very complex evaluation systems that could not be understood easily by common people working on life science. Yu *et al.* (2002) suggested that Fgenesh had best effect for the gene annotation of *Oryza sativa* L. ssp. *indica 93-11*, but they did not make a detailed and exon-level predictive appraisement. Taking the sequences of the 6th chromosome of *Oryza sativa* L. ssp. *japonica* (30.73 Mb, the 5th in length of all 12 rice chromosomes) as analysis data and using steady-going and comparative new edition of Fgenesh (ver. 2.0) as analysis tools, we explored detailedly exon-level gene prediction of monocotyledonous model plant rice and gave a facility evaluation in this study. In order to provide some scientific evidence for genome annotation, especially for gene identification of the food-supplying grass family plants in monocotyledon, further, give a definite target for gene checking and increasing the successful rate of gene identification, we emphasized particularly on testing the gene prediction accuracy by Fgenesh (ver. 2.0) from comparative long, whole sequences of chromosome 6th level, then grouped for the prediction rule of this software by statistical data for high prediction accuracy of exon-type, exon-number of multi-exon genes, and the length distribution of single-exon genes.

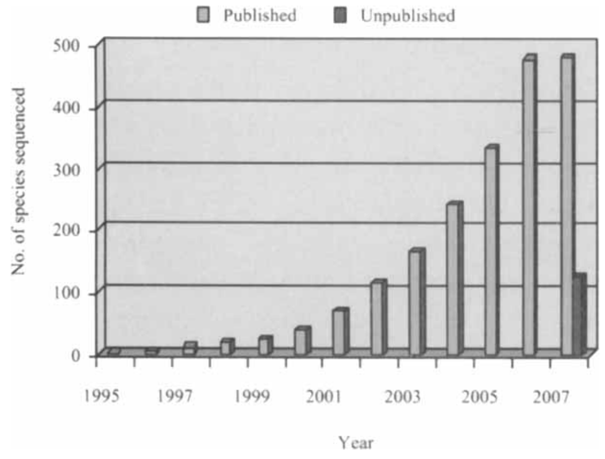

Fig. 1 The statistic of completely sequenced genomes by now.

MATERIALS AND METHODS

Data resource The 6th chromosome sequence, whole genome cDNA (complementary DNA) and EST (expressed sequence tagged) sequence and gene module sequence of *O. sativa* L. ssp. *japonica cv.* Nipponbare were all downloaded from the very famous web site – The Institute for Genomic Research (TIGR, http://www.tigr.org/).

Computational environment SGI3800 server, SUN minitype server, *Lenovo* DeepComp 1800 server, personal computer.

Data disposal process First of all, taking the monocotyledon parameter file, gene predicting for the sequence of the 6th chromosome of *O. sativa* L. ssp. *japonica cv.* Nipponbare was done by Fgenesh (ver. 2.0) running on server. Secondly, by using the tools of small program package developed by perl language, distilled the predicting sequences of FirstExons, InternalExons, LastExons, Introns, SingleExon genes, and non SingleExon genes from the corresponding location of the 6th chromosome sequence. Thirdly, by taking the whole genome cDNA, EST, and gene module sequence as database, the exon and coding sequence mentioned above as query, ran the Blast (basic local alignment search tool) N program on the server [parameter setting was as, matrix: blastn matrix 1-3, gap penalties: (5, 2), expect: 1.0e-10]. Finally, picked the significant (E value < 1e-10) exons and genes from the Blast results by perl programs developed by our research group and made statistical analysis for them.

RESULTS

Prediction results

Fgenesh predicted 4 862 genes for the 6th chromosome