The transfer RNA genes in *Oryza sativa* L. ssp. *indica*

WANG Xiyin (王希胤)[1,2,3*], SHI Xiaoli (史晓黎)[1,2*] & HAO Bailin (郝柏林)[2,4]

1. College of Life Science, Peking University, Beijing 100871, China;
2. Beijing Genomics Institute/Center of Genomics and Bioinformatics, Chinese Academy of Sciences, Beijing 101300, China;
3. Department of Mathematics and Physics, Hebei Institute of Science and Technology, Tangshan 063009, China;
4. The T-life Research Center, Fudan University, Shanghai 200433, China

Correspondence should be addressed to Hao Bailin (email: hao@itp.ac.cn)

Received July 26, 2002

Abstract The availability of the draft genome sequence of *Oryza sativa* L. ssp. *indica* has made it possible to study the rice tRNA genes. A total of 596 tRNA genes, including 3 selenocysteine tRNA genes and one suppressor tRNA gene are identified in 127551 rice contigs. There are 45 species of tRNA genes and the revised wobble hypothesis proposed by Guthrie and Abelson is perfectly obeyed. The relationship between codon usage and the number of corresponding tRNA genes is discussed. Redundancy may exist in the present list of tRNA genes and novel ones may be found in the future. A set of 33 tRNA genes is discovered in the complete chloroplast genome of *Oryza sativa* L. ssp. *indica*. These tRNA genes are identical to those in ssp. *japonica* identified by us independently from the origional annotation.

Keywords: tRNA gene, wobble hypothesis, codon usage, pseudo-tRNA gene, *Oryza sativa*, chloroplast.

Rice is the cereal consumed the most by human, having a history of cultivation more than 8000 years[1]. Great progress has been made recently in rice genome projects[2—4] and a complete genomic map will be available in the near future.

Transfer RNA is an important non-coding RNA involved in protein synthesis, playing a critical role in the fidelity of information transfer between mRNA and protein. Carrying an amino acid molecule, tRNA reads the information contained in mRNA through the interaction of its three-letter anticodon with the three-letter codon of mRNA. There are 61 different encoding codons. However, organisms do not have 61 tRNA species with all possible anticodons. In 1966, Crick proposed the famous “wobble hypothesis”: through non-Watson-Crick base-pairing rules, less tRNA species are needed[5]. Guthrie and Abelson updated and revised the wobble hypothesis in 1982[6]. In the standard genetic code table, in a “two-codon box” two codons ending with U and C encode a different amino acid from the other two codons ending with A and G, in a “four-codon box” all four codons ending with U, C, A and G encode the same amino acid. For example, His and Gln are in a two-codon box and Pro and Gly are in different four-codon boxes. In a two-codon box, the codons ending with U and C are decoded by the anticodon beginning with G (table 1). In a four-codon box, the codons ending with U and C are decoded by the anticodon beginning with A.

* These authors contributed equally to this work.

The only exception is the four-codon box of Gly in which the codons ending with U and C are decoded by the anticodon beginning with G. In the two- and four-codon boxes, the codon ending with A is decoded by the anticodon beginning with U, while the codon ending with G is decoded by the anticodon beginning with C. Based on the modified base-pairing rules, Guthrie and Abelson predicted that in eukaryotes a total of 46 different tRNA species would be enough.

Eukaryotes such as *H. sapiens*, *A. thaliana* and *C. elegans* whose complete collection of tRNA genes is known follow the revised wobble hypothesis almost perfectly, though there are one to three exceptional tRNA genes (table 2)[7,8]. In table 2, we show the amino acids (aa.) in calligraphic letter, followed by the codon, the number of tRNA genes in *A. thaliana, C. elegans, H. sapiens* and the corresponding anticodon (anti). The tRNA gene number of those exceptions to the wobble hypothesis, namely, seven "*1*", is shown in italic. The exceptions may be caused by sequencing errors or they may be pseudogenes. Experimental work is needed to further check

Table 1 The wobble base-pairing rules revised by Guthrie and Abelson for eukaryotes

	Codon (base 3)	Anticodon (base 1)
In any box	A	U[a)]
In any box	G	C
In two-codon box	U&C	G
In four-codon box	U&C	A[b)]
One exceptional four-codon box: Gly	U&C	G

a) U is always modified to other bases such as pseudouridine. b) A is almost certainly modified to I (Isonine).

Table 2 Transfer RNA genes in *A. thaliana* (A), *C. elegans* (C), *H. sapiens* (H)

aa	codon	A	C	H	anti	aa	codon	A	C	H	anti	aa	codon	A	C	H	anti	aa	codon	A	C	H	anti
F	UUU	0	0	0	AAA	S	UCU	37	14	10	AGA	Y	UAU	0	0	*1*	AUA	C	UGU	0	0	0	ACA
	UUC	16	16	14	GAA		UCC	*1*	0	0	GGA		UAC	76	19	11	GUA		UGC	15	13	30	GCA
L	UUA	6	5	8	UAA		UCA	9	7	5	UGA	*	UAA	0	0	1	UUA	*	UGA	0	0	0	UCA
	UUG	10	7	6	CAA		UCG	4	5	4	CGA		UAG	0	0	1	CUA	W	UGG	14	11	7	CCA
	CUU	11	18	13	AAG	P	CCU	16	6	11	AGG	H	CAU	0	0	0	AUG	R	CGU	9	18	9	ACG
	CUC	*1*	0	0	GAG		CCC	0	0	0	GGG		CAC	10	17	12	GUG		CGC	0	*1*	0	GCG
	CUA	10	3	2	UAG		CCA	39	34	10	UGG	Q	CAA	8	18	11	UUG		CGA	6	10	7	UCG
	CUG	3	5	6	CAG		CCG	5	3	4	CGG		CAG	9	7	21	CUG		CGG	4	3	5	CCG
I	AUU	20	19	13	AAU	T	ACU	10	17	8	AGU	N	AAU	0	0	*1*	AUU	S	AGU	0	0	0	ACU
	AUC	0	0	*1*	GAU		ACC	0	0	0	GGU		AAC	16	20	33	GUU		AGC	13	9	7	GCU
	AUA	5	8	5	UAU		ACA	8	11	10	UGU	K	AAA	13	16	16	UUU	R	AGA	9	7	5	UCU
M	AUG	23	20	17	CAU		ACG	6	7	7	CGU		AAG	18	33	22	CUU		AGG	8	3	4	CCU
V	GUU	15	19	20	AAC	A	GCU	16	21	25	AGC	D	GAU	0	0	0	AUC	G	GGU	*1*	0	0	ACC
	GUC	0	0	0	GAC		GCC	0	0	0	GGC		GAC	23	22	10	GUC		GGC	23	14	11	GCC
	GUA	7	6	5	UAC		GCA	10	10	10	UGC	E	GAA	12	17	14	UUC		GGA	12	33	5	UCC
	GUG	8	5	19	CAC		GCG	7	4	5	CGC		GAG	13	20	8	CUC		GGG	5	3	8	CCC

whether they are genuine tRNA genes. Therefore, it is interesting to count how many tRNA genes are there in rice and whether the revised wobble hypothesis holds.

Some codons are used more frequently than others. As an important characteristic of a genome, codon usage is quite instructive for gene-finding. In a genome, some codons are preferred to other synonymous ones. This phenomenon is called codon bias. All organisms show codon bias. Highly expressed genes show the strongest codon bias in less complex organisms[7]. The relationship between codon usage and tRNA genes in rice genome will be discussed below.

Some special tRNA genes are also dealt with in the present work. Pseudo-tRNA genes are those genomic sequences structurally related to genuine tRNA genes but functionally inactive because of insertions, deletions or lacking of functional promoters. Experimental work is usually required to check whether those predicted pseudogenes are inactive or not. Suppressor tRNA is a mutant tRNA that recognizes a nonsense codon (UAA/UAG) instead of the codon for the cognate amino acid. The mutation is, sometimes but not always, caused by a base substitution in the anticodon. Selenocysteine tRNA gene is a type of tRNA carrying selenocysteine, which is encoded by the UGA codon (a nonsense codon) during protein synthesis.

There are two cultivated subspecies of rice, *indica* and *japonica*. The chloroplast genome sequences of both the subspecies are now available. Transfer tRNA genes are identified and compared in the two chloroplast genome sequences.

1 Material and methods

All 127551 contigs of *indica* were retrieved from the Rice GD[9]. The chloroplast genome sequence of *indica* was sequenced by Beijing Genomics Institute. The chloroplast genome sequence of *japonica*[10] was retrieved from the GenBank[11].

Three public programs were used. The software tRNAscan-SE[12, 13] was first run on those contigs and on both chloroplast genomes to discover possible candidate tRNA genes. In order to improve the reliability of the result these candidate tRNA genes were taken as query sequences to run BLASTN[14] search against all sequences in the non-redundant (nr) database at CBI[15]. Multi-alignment was performed by using the software ClustalW[16].

2 Results and discussion

2.1 The classification of the canonical tRNA genes

With the help of tRNAscan-SE, we discovered 881 candidate tRNA genes in rice contigs. On the basis of scores and identities during BLASTN, a total of 592 canonical tRNA genes were further confirmed and classified into three groups: BLASTN-confirmed tRNA genes with a BLASTN score larger than 115, probable novel tRNA genes with a score smaller than 115 but larger than 100, putative novel tRNA genes with a score smaller than 100 but larger than 80.

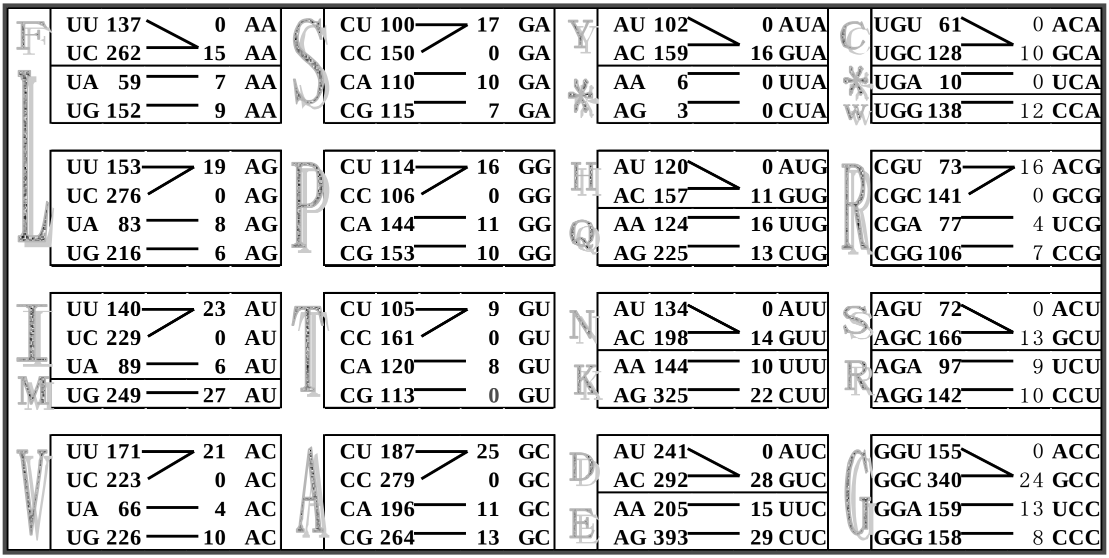
There are 467 BLASTN confirmed tRNA genes which are the most authentic tRNA genes. Without perfect match with other eukaryotic tRNA genes in the nr database, most of the 74 probable novel tRNA genes and the 51 putative novel tRNA genes are considered to be novel tRNA genes adapted to rice.

A set of 596 tRNA genes in *japonica* is listed in the TIGR DB[17]. The tRNA gene number in *japonica* is quite close to that in *indica*.

2.2 The wobble hypothesis

The whole set of canonical tRNA genes in rice include 45 species and the revised wobble hypothesis proposed by Guthrie and Abelson is perfectly followed (table 3) with only one exception for the time being (see sec. 2.7 below). In table 3, the 20 amino acids are denoted in calligraphic letters. The table shows the codons, codon frequency per 10000 codons, the number of tRNA genes and corresponding anticodons. Base-pairing rules are indicated by black lines. The codon usage data are derived from the Codon Usage Database[18] in which 27910 codons of *indica* were counted.

Table 3 The rice codons and the associated tRNA genes



F	UU	137	0	AA	S	CU	100	17	GA	Y	AU	102	0	AUA	C	UGU	61	0	ACA
	UC	262	15	AA		CC	150	0	GA		AC	159	16	GUA		UGC	128	10	GCA
L	UA	59	7	AA		CA	110	10	GA	*	AA	6	0	UUA	*	UGA	10	0	UCA
	UG	152	9	AA		CG	115	7	GA		AG	3	0	CUA	W	UGG	138	12	CCA
	UU	153	19	AG	P	CU	114	16	GG	H	AU	120	0	AUG	R	CGU	73	16	ACG
	UC	276	0	AG		CC	106	0	GG		AC	157	11	GUG		CGC	141	0	GCG
	UA	83	8	AG		CA	144	11	GG	Q	AA	124	16	UUG		CGA	77	4	UCG
	UG	216	6	AG		CG	153	10	GG		AG	225	13	CUG		CGG	106	7	CCG
I	UU	140	23	AU	T	CU	105	9	GU	N	AU	134	0	AUU	S	AGU	72	0	ACU
	UC	229	0	AU		CC	161	0	GU		AC	198	14	GUU		AGC	166	13	GCU
	UA	89	6	AU		CA	120	8	GU	K	AA	144	10	UUU	R	AGA	97	9	UCU
M	UG	249	27	AU		CG	113	0	GU		AG	325	22	CUU		AGG	142	10	CCU
V	UU	171	21	AC	A	CU	187	25	GC	D	AU	241	0	AUC	G	GGU	155	0	ACC
	UC	223	0	AC		CC	279	0	GC		AC	292	28	GUC		GGC	340	24	GCC
	UA	66	4	AC		CA	196	11	GC	E	AA	205	15	UUC		GGA	159	13	UCC
	UG	226	10	AC		CG	264	13	GC		AG	393	29	CUC		GGG	158	8	CCC

2.3 Codon usage and tRNA genes

We studied codon bias as well as the correlation of codon usage with the corresponding tRNA gene copy number (table 3). The codons of GAG (393), GGC (240), AAG (325), GAC (292) and GCC (279) are much more frequently used than those of CUA (83), CGA (77), GUA (66), UGU (61) and UUA (59). Rice prefers codons ending with C or G to those ending with U or A. We checked all 16 groups of codons, the codons in each of the groups having the same first two letters. Without counting the terminal codons and UGG encoding Trp, we see that in 15 groups the usage of codons ending with C exceeds that ending with U whereas in 14 groups the usage of codons ending with G predominates that ending with A. Two pairs of exceptions are (CCU, CCC) and (ACA, ACG).

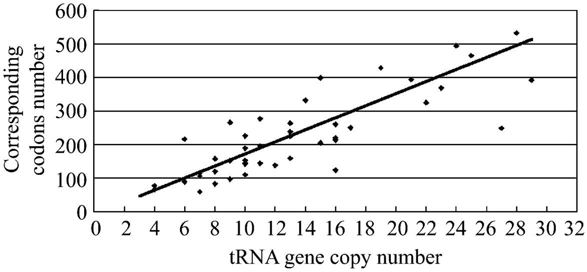


Fig. 1. The correlation of tRNA gene numbers with the corresponding codon numbers.

There exists a rough positive correlation between codon usage and the corresponding tRNA gene copy number in rice (fig. 1) as most organisms do.

We calculated the ratio of codon frequencies to the corresponding tRNA gene numbers for all the codons classified by the third base (table 4). The ratio for codons ending with U and C is larger than that for codons ending with A or G. One possible explanation is that those tRNA genes related to wobble codons are transcriptionally more efficient or that the corresponding tRNAs are more functional. Similarly, the tRNA genes related to codons ending with G are more efficient than those related to codons ending with A in rice.

Table 4 On the transcriptional or functional efficiency of tRNA genes

	Codon frequency	Corresponding tRNA gene number	Codon frequency per tRNA gene
Codons ending with U and C	5332	277	19.25
Codons ending with A	1678	132	12.67
Codons ending with G	2978	183	16.26

The data were derived from table 3.

2.4 Pseudo-tRNA gene

We discovered 27 pseudo-tRNA genes in the rice contigs. With the help of tRNAscan-SE, which also searches possible pseudo-tRNA genes structurally related to genuine tRNA genes, a group of 71 pseudo-tRNA genes were found. However, most of them were discarded after doing BLASTN for having no similarity to authentic eukaryotic tRNA genes or being similar to bacteria, chloroplast or mitochondria sequences. The pseudo-tRNA genes confirmed have some similarity to the eukaryotic tRNA genes or high similarity to known rice sequences in database. The anticodons of more than half of the pseudo-tRNA genes could not be clearly determined.

It is impossible to decide by computational analysis alone whether these are functional genes or pseudogenes. Further experimental study needs to be conducted.

2.5 Selenocysteine tRNA gene and suppressor tRNA genes

There are 3 selenocysteine tRNA genes and one suppressor tRNA gene. Selenocysteine is sometimes called the 21st encoded natural amino acid found in every domain of life on Earth (the 22nd natural amino acid pyrrolysine encoded by UAG has been found recently[19,20]). In fact, selenocysteine is encoded by the umber (UGA) codon. Different mechanisms are adopted in prokaryotes and eukaryotes to tell the translational machinery of the cells whether it should continue or terminate the process of translation. The only possible suppressor tRNA gene transcribes a tRNA recognizing the ochre (UAA) codon.

2.6 Intron in tRNA genes

Making up to 6% of the total, there are 36 tRNA genes with an intron identified by tRNAscan-SE. All of them have only one intron with an average length of 15 bases. The longest intron in the trnS-AGA gene is 38 bases long. All trnY-GUA genes have an intron. All of the elongation trnM genes have an intron, while all of the initiator trnM genes have no intron. The only possible suppressor tRNA-UUA gene has a 21-base intron.

2.7 The absence of trnT-CGU gene

The trnT-CGU genes are absent in table 3. In fact, six possible trnT-CGU genes were found by tRNAscan-SE but discarded for low similarity to the known tRNA genes. Since there are only 356M bases in the assembled *indica* contigs, a significant fraction of the rice genome whose estimated euchromatic size is 464 Mb is not available yet in the present collection of contigs. The absence of trnT-CGU genes might be caused by the incompleteness of the genome contigs. The situation might be enhanced by the property of tRNA genes' tending to cluster together in a chromosome. This can be illustrated by that a group of 8 trnQ-UUG genes were found on a contig in rice and 140 tRNA genes, making up to 25% of the total, cluster in a narrow region of only 4M on chromosome 6 in human[7].

2.8 Comparison of tRNA genes among eukaryotes

The numbers of tRNA genes of seven eukaryotes are listed in table 5. The ratio of tRNA gene numbers to the overall size of the genomes decreases drastically with the increase of genome size, while the ratio to the total size of coding sequences (CDS) decreases gradually. This might hint that the efficiency of tRNA gene is greater in higher organisms.

Table 5 Transfer RNA genes in eukaryotes

Species	tRNA gene number	Genome size /Mbp	tRNA gene number per Mbp in genome	CDS size /Mbp	tRNA gene number per Mbp of CDS
S. cerevisiae	273	12	22.75	8.45	32
S. pombe	174	14	12.43	6.9	25
C. elegans	584	100	5.84	26.1	22
A. thaliana	620	125	4.96	33.5	18
D. melanogaster	284	180	1.58	24.1	12
O. sativa	596	464	1.48	–	–
H. sapiens	648	3400	0.19	58.5 (?)	11 (?)

The data in this table were partly derived from refs. [8, 11, 21, 22].

The transfer RNA genes in *A. thaliana* are more conserved with fewer mutations, insertions and deletions, while those in rice and human have more inter-subspecies mutations. About 17 subspecies tRNA genes in rice, making up to one fourth of the total subspecies, have perfectly conserved counterparts in *Arabidopsis*. None tRNA gene in human has an identical counterpart in rice or *Arabidopsis*. The trnP-AGG, trnP-CGG, trnA-AGC, trnA-CGC genes are the most conserved tRNA genes, while the trnE-CTC and trnK-TTT genes are the least conserved in rice, human and *Arabidopsis*.

2.9 The comparative analysis of tRNA genes arrangement between rice and *Arabidopsis*

Since we have not located the contigs of *indica* in chromosomes so far, it is impossible to perform the whole genome tRNA genes arrangement comparison between rice and *Arabidopsis*. However, tRNA genes that lie in one contig can reveal local arrangement of tRNA genes in chromosomes. All ten gene clusters containing more than two tRNA genes of *indica* were focused on and all members of each cluster are in the same contig. We tried to find the analogous structure of arrangement in *Arabidopsis* for those ten rice gene clusters. The largest rice gene cluster has eight trnQ-TTG. Each gene in the cluster is 71 bases. Those genes are separated by approximate two thousands bases. However only one chromosome of *Arabidopsis* has more than two trnQ-TTG, that is, four trnQ-TTG were identified by tRNAscan-SE on chromosome 1 of *Arabidopsis*. The average distance of these genes is more than eight millions bases. Also no analogous tRNA genes arrangement structure was found in the genome of *Arabidopsis* for the other nine clusters of rice tRNA genes.

There are 27 three-gene clusters with trnS-AGA, trnY-GTA and trnY-GTA in fixed order in the *Arabidopsis* chromosome 1. Each cluster is about hundreds bases in length. We did not find any similar gene structure in the rice contigs. Our current research indicates that the arrangement of tRNA genes in chromosomes of rice and *Arabidopsis* has undergone huge change after the division of dicot and monocot.

2.10 Chloroplast tRNA genes in *japonica* and *indica*

The chloroplast tRNA genes in the two cultivated rice subspecies, *indica* and *japonica*, were also studied. We found in the *indica* chloroplast genome 33 tRNA genes using tRNAscan-SE with subsequent BLASTN search. We also found 33 tRNA genes in the chloroplast genome of the *japonica* subspecies in the same way. We note that 18 tRNA genes from the 33 tRNA genes of the latter set are different from those given in the annotation of the GenBank entry[14,15]. Multi-alignment by using the software ClustalW shows that the tRNA genes in the two subspecies are perfectly conserved. It is a remarkable fact that in spite of more than 7000 years of separation no mutation could be observed in the chloroplast tRNA genes in the two subspecies.

We note in conclusion that the strategy of using tRNAscan-SE in combination with BLASTN searching for tRNA genes is efficient. However, this set of rice tRNA genes may be redundant and novel tRNA genes may be found in the future. Experimental work is needed to check whether they are genuine tRNA genes.

Acknowledgements We thank Ni Peixiang and all the participants of the Rice Genome Sequencing Project at the Beijing Genomics Institute/Center of Genomics and Bioinformatics, Chinese Academy of Sciences. We gratefully acknowledge the support of the Center of Bioinformatics at Peking University, where part of this research was carried out. Hao acknowledges the support from the special Funds for Major State Basic Research Projects and the Beijing Municipality “248” Innovation Project.

References

1. Zhang, J. H., Wang, X. K., Kong, Z. C., Rice cultivation of Jiahu Remains in Henan Province, Science J. (in Chinese), 2002, 54(3): 3.
2. Yu, J., Hu, S. N., Wang, J. et al., A draft sequence of the rice (*Oryza sativa* L. ssp. *indica*) genome, Chinese Science Bulletin, 2001, 46(23): 1937—1941.
3. Yu, J., Hu, S. N., Wang, J. et al., A draft sequence of the rice genome (*Oryza sativa* L. ssp. *indica*), Science, 2002, 296: 79.
4. Goff, S. A., Ricke, D., Lan, T. H. et al., A draft sequence of the rice genome (*Oryza sativa* L. ssp. *japonica*), Science, 2002, 296: 92.
5. Crick, F., Codon-anticodon pairings: the wobble hypothesis, Journal of Molecular Biology, 1966, 19: 548.
6. Guthrie, C., Abelson J., Organization and expression of tRNA genes in *Saccharomyces cerevisiae*, in The Molecular Biology of the Yeast *Saccharomyces*: Metabolism and Gene Expression (eds. Strathern J., et al.), New York: Cold Spring Harbor Laboratory Press, 1982, 487.
7. International Human Genome Sequencing Consortium, Initial sequencing and analysis of the human genome, Nature, 2002, 409: 860.
8. http://rna.wustl.edu/GtRDB/
9. Rice GD at Beijing Genomics Institute: http://btn.genomics.org.cn/rice
10. Hiratsuka, J., Shimada, H., Whittier, R. et al., The complete sequence of the rice (*Oryza sativa*) chloroplast genome: intermolecular recombination between distinct tRNA genes accounts for a major plastid DNA inversion during the evolution of the cereals, Molecular General Genetics, 1989, 217 (2-3): 185.
11. http://www.ncbi.nlm.nih.gov/
12. Eddy, S. R., Non-coding RNA genes and the modern RNA world, Nature Reviews Genetics, 2001, 2: 919.
13. Lowe, T. M., Eddy S. R., tRNAscan-SE: A program for improved detection of transfer RNA genes in genomic sequence, Nucleic Acids Research, 1997, 25: 955.
14. Altschul, S. F., Madden, T. L., Shaffer, A. A. et al., Gapped BLAST and PSI-BLAST: A new generation of protein database search programs, Nucleic Acids Research, 1997, 25: 3389.
15. Center for Bioinformatics, Peking University: http://www.cbi.pku.edu.cn/
16. Thompson, J. D. et al., CLUSTALW: improving the sensitivity of progressive multiple sequence alignment through sequence weighting, position-specific gap penalties and weight matrix choice, Nucleic Acids Research, 1994, 22: 4673.
17. http://www.tigr.org/
18. http://www.kazusa.or.jp/codon/
19. Srinivasan, G., James, C. M., Krzycki, J. A. et al., Pyrrolysine encoded by UAG in archaea: Charging of a UAG-decoding specialized tRNA, Science, 2002, 296: 1459.
20. Hao, B., Gong, W. M., Ferguson, T. K. et al., A new UAG-encoded residue in the structure of a methanogen methyltransferase, Science, 2002, 296: 1462.
21. Adams, M. D., Celniker, S. E., Holt, R. A. et al., The genome sequence of *Drosophila melanogaster*, Science, 2000, 287: 2185.
22. Goffeau, A., Barrell, B. G., Bussey, H. et al., Life with 6000 genes, Science, 1996, 274: 546.