Journal of Zhejiang University SCIENCE
ISSN 1009-3095
http://www.zju.edu.cn/jzus
E-mail: jzus@zju.edu.cn

Maximal sequence length of exact match between members from a gene family during early evolution*

WEN Xiao (温 晓)[1], GUO Xing-yi (郭兴益)[2], FAN Long-jiang (樊龙江)[†1,2]

([1]*Institute of Crop Science, Zhejiang University, Hangzhou 310029, China*)

([2]*Institute of Bioinformatics/IBM Biocomputational Lab, Zhejiang University, Hangzhou 310029, China*)

[†]E-mail: fanlj@zju.edu.cn

Received Nov. 16, 2004; revision accepted Feb. 19, 2005

Abstract: Mutation (substitution, deletion, insertion, etc.) in nucleotide acid causes the maximal sequence lengths of exact match (MALE) between paralogous members from a duplicate event to become shorter during evolution. In this work, MALE changes between members of 26 gene families from four representative species (*Arabidopsis thaliana*, *Oryza sativa*, *Mus musculus* and *Homo sapiens*) were investigated. Comparative study of paralogous' MALE and amino acid substitution rate (d_A<0.5) indicated that a close relationship existed between them. The results suggested that MALE could be a sound evolutionary scale for the divergent time for paralogous genes during their early evolution. A reference table between MALE and divergent time for the four species was set up, which would be useful widely, for large-scale genome alignment and comparison. As an example, detection of large-scale duplication events of rice genome based on the table was illustrated.

Key words: Maximal length of exact match (MALE), Divergent time, Gene family, Minimal length of exact match (MILE), Genome alignment
doi:10.1631/jzus.2005.B0470 **Document code:** A **CLC number:** Q81

INTRODUCTION

Gene duplication provides a main resource of new genes in genomes (Ohno, 1970; Brown, 1999). Sequences of two paralogous genes from a duplication event will become different from each other along with evolutionary processes. And the difference in sequences caused by substitution, deletion, insertion of nucleotide acid will cause maximal sequence lengths of exact match (MALE) between paralogous members from a gene family to become shorter during evolution. As for example, of a newborn gene, its sequence is the same as that of its parental (paralogous) gene, i.e. the MALE equals to the length of the original sequence. During the evolutionary process, the MALE would become shorter gradually when genetic mutations occurred. As for gene families, the exact trend of the MALE changes among paralogous genes still remain obscure, and a detailed description of the relationship between MALE and evolutionary time is needed. For random sequences, the longest match and its statistical significance between two random DNA sequences has been well documented in early study on modelling a random DNA sequence alignment (Arratia *et al.*, 1986; Karlin and Altschul, 1990).

In large-scale genome sequence alignment, a parameter named minimal sequence length of exact match (MILE), was usually used in corresponding algorithms to limit the searching space and return their results in practical time, such as algorithms implicated by MUMmer, a very fast and widely used program for large-scale genome alignment and comparison (Delcher *et al.*, 2002). In such algorithms, all exact matches between two target genomic se-

* Project supported by the National Natural Science Foundation of China (Grant Nos. 30270810, 90208022 and 30471067) and IBM Shared University Research (Life Science), China

quences will be located at the first step. All exact matches under the value of MILE which was set beforehand, will be ignored in the next alignment. Gene sequences (coding sequences or translated amino acid sequences) on the genomic sequences are usually used in genome alignment such as PROmer of MUMmer packet, which translates target genomic sequence into protein sequences through six open reading frames. In practice, the value of MILE is usually regulated (default value was set as 6 aa in PROmer) for rational hits to be returned. High MILE value chosen, leads to fewer hits returned. Apparently, to gene sequences, MILE implicates some kinds of similarity, or evolutionary scale. But the evolutionary significance of MILE in genome alignment has not been reported yet.

In this study, we investigated the MALE changes of main gene families from four representative species (*Arabidopsis thaliana*, *Oryza sativa*, *Mus musculus* and *Homo sapiens*) during evolution. Our results indicated that MALEs were related with evolutionary time significantly during early evolution of gene families, and four curve functions between MALE and evolutionary time were created for the four species using data on their gene families, respectively. Based on the four functions, a reference table between MALE and its corresponding evolutionary time was also constructed for the four species. As an important application of our study on genome alignment, an example of study on large-scale duplication events of rice genome was illustrated.

MATERIALS AND METHODS

Sequence data source

Protein sequences of 9 gene families (GTPBP, SCDehydRed, MFS, UDPGlycTnsf, GSDLLipase, Polygalns, SubtilisinSP, CytP450 and Calmod) of *Arabidopsis thaliana* were downloaded from http://www.tc.umn.edu/~cann0010/genefamilyevolution/index.html and 17 gene families of other three species (LTP, Peroxidase, ABC, NbsLrr and CytP450 of *Oryza sativa*; Mage, Hox, ABC, Potassium voltage-gated channel, Matrix metalloproteinase, kallikrein and CytP450 of *Homo sapiens*; ABC, Collagen, Hox, Synaptotagmin and CytP450 of *Mus musculus*) were selected from the protein database of NCBI (http://www.ncbi.nlm.nih.gov). Average number of members of a gene family was 98 (ranging from 92 to 112).

Gene Family Criteria: Amino acid sequences of members from gene families must have over 40% sequence similarity and contain all key amino acid signature motifs of the corresponding gene family. Two hundred random sequences were created using PERL program.

Estimation of MALE and amino acid substitution rates of paralogous genes

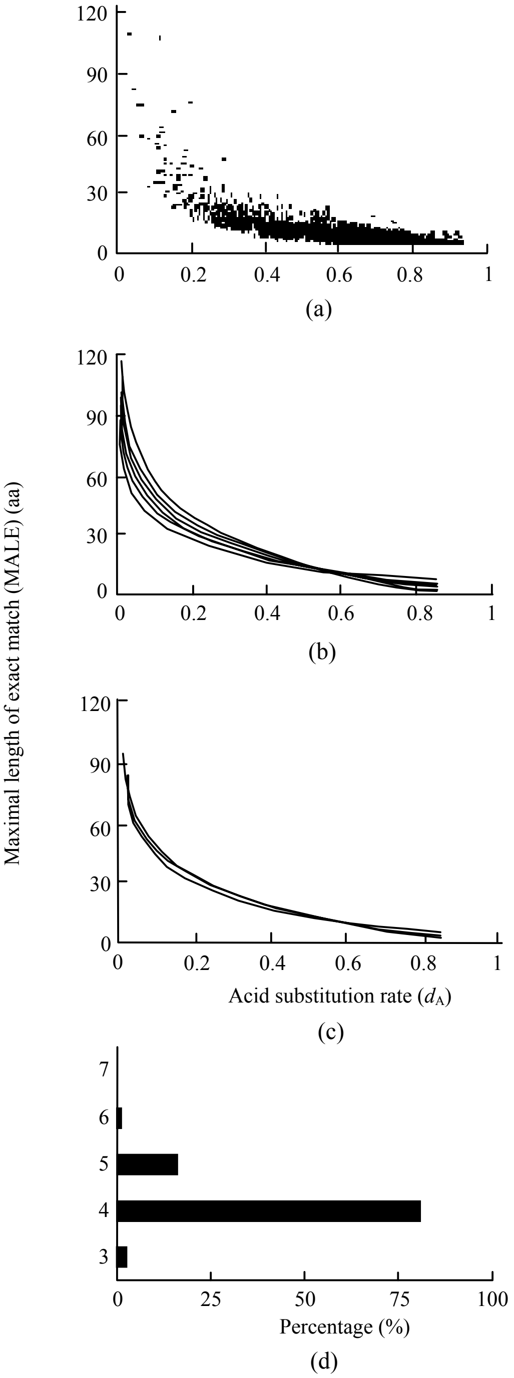
MALEs between paralogous genes from a gene family were calculated using mummer program of MUMmer package (Delcher *et al*., 2002). The amino acid substitution rate (d_A) was estimated using the aaml program of PAML package (Yang, 1999) with the Dayhoff matrix. The divergence time was calculated based on a molecular clock rate of 9×10^{-10} nonsynonymous substitutions per site per lineage per year and 2.25 nonsynonymous substitutions per amino acid change (Lynch and Conery, 2000; Goff *et al*., 2002).

RESULTS

MALE changes during evolution

In order to show an enlarged picture of MALE changes of gene families during evolution, those big size gene families from four representative species were chosen. A total of 26 gene families averaging 98 members per family were used in this study. For brevity, we depicted only one big gene family, cytochrome P450 (Fig.1). The other families, however, did behave as we had expected, and consistent with their evolutionary clade.

A clear changing trend between MALE and amino acid substitution rate (d_A) of paralogs from a gene family could be observed in the dot-plot of the *Arabidopsis* P450 gene family for low d_A values (<0.5) (Fig.1a), and therefore, a curve function could be created to fit it significantly (Fig.1c). The results indicated that MALEs of paralogous genes decreased sharply at the beginning period of evolution (d_A<0.1) and then became slower during early evolution (d_A<0.5). Along with the further accumulation of mutations (d_A>0.5), MALE decreased very slowly and



(a)

(b)

(c)

(d)

Fig.1 Changes of maximal length of exact match (MALE) among paralogous genes from a gene family during evolution (a) dot-plot of cytochrome P450 gene family from *Arabidopsis;* (b) MALE as function of amino acid substitution rate (d_A). Shown here are 9 gene families (GTPBP, SCDehydRed, MFS, UDPGlycTnsf, GSDLLipase, Polygalns, SubtilisinSP, CytP450 and Calmod) from *Arabidopsis*; (c) also MALE as function of amino acid substitution rate (d_A). Shown here are 4 cytochrome P450 gene families from four species (*Arabidopsis*, *Oryza sativa*, *Mus musculus* and *Homo sapiens*), respectively; (d) frequency distribution of MALEs between random sequences

ultimately stayed around a fixed value (6–10 aa). MALEs distribution of random sequences indicated that almost (95%) of their MALEs fell into a range of 3–6 aa (Fig.1d). Besides the random sequences by computer simulation, we also investigated the MALE using real biological sequences by random by choosing 93 protein sequences of 27 species from Swiss-Prot protein database. The results were similar to those by computer simulation, and suggested that MALEs were related with divergence time significantly during early evolution of the gene families.

The trend of MALE changes of P450 gene family presented almost the same pattern in four different species (Fig.1c), representing a wide range of the biological kingdom (*Arabidopsis* and *Oryza sativa* for dicot and monocot plants, *Homo sapiens* and *Mus musculus* for vertebrate). The results suggested that MALEs of a common gene family tended to behave similarly during evolution in different species. Then how did the different gene families in the same species behave? Based on nine gene families from *Arabidopsis*, the results showed that MALE changes (lines) of different gene families were basically the same (Fig.1b), although there existed some differences among them, which actually stemmed from the different evolutionary speed among genes and gene families in a genome.

MALE and divergent time

The above results indicated that MALE changes could be considered as a good scale of evolutionary time during early evolution of gene families, and different gene families from the same species basically present a common pattern of MALE changes during evolution. This pointed out the possibility of setting up a reference table relating MALE and evolutionary time at a species (genome) level. The table would be useful and convenient for sequence analysis at genome level, such as large-scale genome alignment. In such a table (Table 1), MALEs and corresponding d_A values and divergence times were averaged from representative gene families of the four species. Therefore, duplicate events that occurred in a genome at different times during evolution can be estimated by just choosing different MALE values based on the table. A detailed example will be illustrated in the next section.

Reference tables for four species were considered

Table 1 Reference table between maximal sequence lengths of exact match (MALE) and divergent time. Average amino acid substitution rates (d_A) and divergent time (MY, million years ago; figures in bracket are plus values of mean and two standard deviations) of four species (*Arabidopsis thaliana*, *Oryza sativa*, *Mus musculus* and *Homo sapiens*) are listed

MALE (aa)	*Arabidopsis*		*Oryza*		*Mus*		*Homo*	
	MY	d_A	MY	d_A	MY	d_A	MY	d_A
7	164 (191)	0.6576	145 (177)	0.5807	168 (189)	0.6701	165 (190)	0.6592
8	158 (185)	0.6332	142 (173)	0.5671	163 (186)	0.6526	162 (187)	0.6462
9	152 (179)	0.6097	138 (169)	0.5539	159 (182)	0.6355	158 (185)	0.6335
10	147 (173)	0.5871	135 (166)	0.5410	155 (179)	0.6189	155 (182)	0.6210
11	141 (167)	0.5653	132 (162)	0.5284	151 (176)	0.6027	152 (179)	0.6087
12	136 (161)	0.5444	129 (158)	0.5161	147 (173)	0.5870	149 (177)	0.5967
13	131 (156)	0.5242	126 (155)	0.5040	143 (170)	0.5716	146 (174)	0.5850
14	126 (151)	0.5047	123 (151)	0.4923	139 (167)	0.5567	143 (172)	0.5735
15	121 (146)	0.4860	120 (148)	0.4808	136 (164)	0.5421	141 (169)	0.5622
16	117 (141)	0.4680	117 (145)	0.4696	132 (162)	0.5279	138 (167)	0.5511
17	113 (136)	0.4506	115 (141)	0.4586	129 (159)	0.5141	135 (165)	0.5402
18	108 (132)	0.4339	112 (138)	0.4479	125 (156)	0.5007	132 (162)	0.5296
19	104 (127)	0.4178	109 (135)	0.4375	122 (153)	0.4876	130 (160)	0.5191
20	101 (123)	0.4023	107 (132)	0.4273	119 (151)	0.4748	127 (158)	0.5089
21	97 (119)	0.3874	104 (129)	0.4173	116 (148)	0.4624	125 (155)	0.4989
22	93 (115)	0.3730	102 (126)	0.4076	113 (145)	0.4503	122 (153)	0.4891
23	90 (111)	0.3592	100 (123)	0.3981	110 (143)	0.4385	120 (151)	0.4794
24	86 (107)	0.3458	97 (121)	0.3888	107 (141)	0.4271	117 (149)	0.4700
25	83(104)	0.3330	95 (118)	0.3797	104 (138)	0.4159	115 (147)	0.4607
26	80(100)	0.3207	93 (115)	0.3709	101 (136)	0.4050	113 (145)	0.4516
27	77(97)	0.3088	91 (113)	0.3622	99 (133)	0.3944	111 (142)	0.4427
28	74 (94)	0.2973	88 (110)	0.3538	96 (131)	0.3841	109 (140)	0.4340
29	72 (90)	0.2863	86 (108)	0.3455	94 (129)	0.3741	106 (138)	0.4255
30	69 (87)	0.2757	84 (105)	0.3375	91 (127)	0.3643	104 (136)	0.4171
35	57 (74)	0.2282	75 (94)	0.2999	80 (116)	0.3191	94 (127)	0.3776
40	47 (62)	0.1889	67 (84)	0.2665	70 (106)	0.2795	85 (118)	0.3418
45	39 (53)	0.1564	59 (75)	0.2368	61 (98)	0.2448	77 (110)	0.3094
50	32 (44)	0.1294	53 (67)	0.2105	54 (89)	0.2144	70 (102)	0.2801
55	27 (37)	0.1071	47 (60)	0.1871	47 (82)	0.1878	63 (95)	0.2536
60	22 (32)	0.0887	42 (53)	0.1662	41 (75)	0.1645	57 (89)	0.2296
65	18 (27)	0.0734	37 (48)	0.1477	36 (69)	0.1441	52 (82)	0.2078
70	15 (22)	0.0608	33 (42)	0.1313	32 (63)	0.1262	47 (77)	0.1882
75	13 (19)	0.0503	29 (38)	0.1167	28 (58)	0.1105	43 (71)	0.1703
80	10 (16)	0.0416	26 (34)	0.1037	24 (53)	0.0968	39 (66)	0.1542
85	9 (13)	0.0345	23 (30)	0.0922	21 (49)	0.0848	35 (62)	0.1396
90	7 (11)	0.0285	20 (27)	0.0819	19 (45)	0.0743	32 (58)	0.1264
95	6 (10)	0.0236	18 (24)	0.0728	16 (41)	0.0651	29 (54)	0.1144
100	5 (8)	0.0196	16 (21)	0.0647	14 (37)	0.0570	26 (50)	0.1036

in this study (Table 1). Based on datasets of their representative gene families, the logarithm equations between MALEs and their d_A values of the four species were constructed: y=−41.11ln(x)−0.11 (R^2=0.92) (*Arabidopsis*), y=−39.79ln(x)−11.92 (R^2=0.94) (*Oryza sativa*), y=−34.14ln(x)−5.45 (R^2=0.91) (*Mus musculus*) and y=−44.54ln(x)−10.36 (R^2=0.90) (*Homo sapiens*). In Table 1, two divergent (duplication) times, mean and maximal time (plus value of mean and two standard deviations) and their corresponding MALE are listed for every species. Maximal time of a MALE means that duplicate events of most (>95%) gene families with the MALE occurred before the time.

Application of MALE

An important objective of MALE application is the analysis of large-scale genome alignment and comparison. Here we displayed an example to illustrate how to detect large-scale duplication blocks in rice genome using above reference table.

In large-scale genome sequence alignment, the minimal sequence length of exact match (MILE) was usually used in corresponding algorithms to limit searching space and return their results in practical time. When a particular MILE was chosen, it means that those duplicate genes with less sequence than MALE, which equates to the MILE, would be ignored and not be returned. For example, when MILE parameter is set at 30 aa in a genome alignment using program PROmer of MUMmer packet, it means only those genes would be focused on that originated within an evolutionary time, say, 84 million years for rice (Table 1), and other genes will be absolutely screened out.

Fig.2 showed the selected inter-chromosome alignment results (between chromosome 1 and 5; 11 and 12) of rice using PROmer program of MUMmer packet. Two levels (30 and 100 aa) of MILE were chosen, which mean those duplicate genes with less than 30 or 100 MALE will be ignored and not be returned. Apparently, there were changes in the dot-

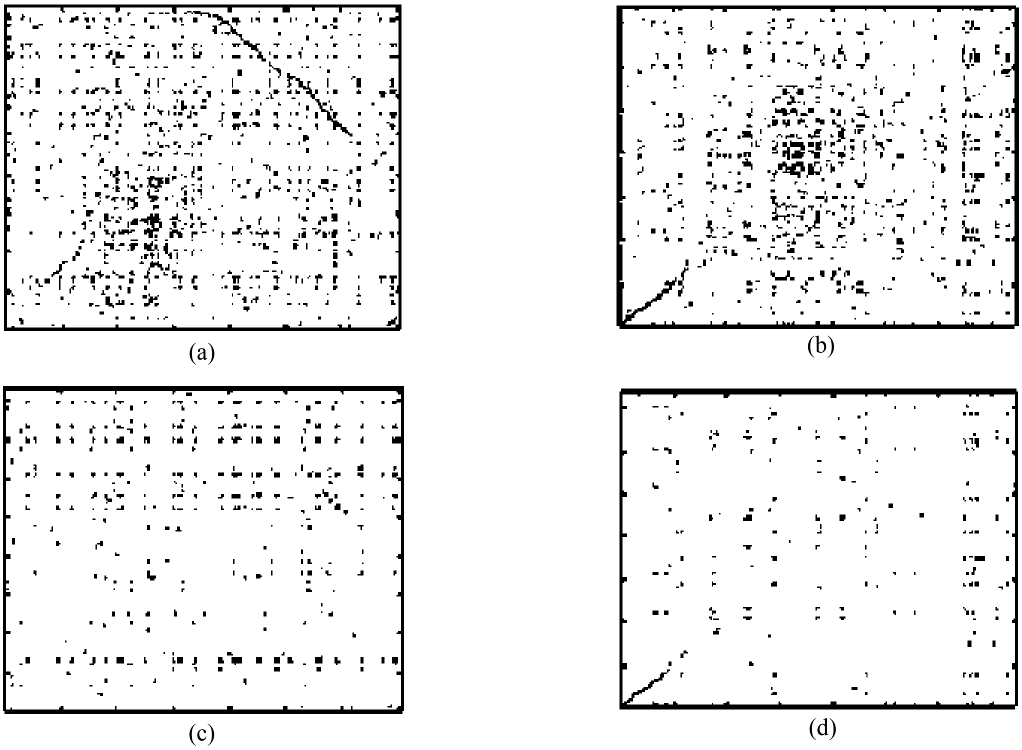


Fig.2 Selected results of large-scale genome alignment of rice. TIGR's rice 12 chromosome pseudomolecules (Version 1.0) and PROmer program of MUMmer package were used (a) Alignment between chromosome 1 and chromosome 5, MILE was limited at 30 aa; (b) Alignment between chromosome 11 and chromosome 12, MILE was limited at 30 aa; (c) Alignment between chromosome 1 and chromosome 5, MILE was limited at 100 aa; (d) Alignment between chromosome 11 and chromosome 12, MILE was limited at 100 aa

plots between chromosome 1 and 5 using the two different MILE parameters: a syntenic line was observed when 30 MILE was chosen and it vanished as MILE increased to 100. The situation of chromosome 11 and 12 was different from that of chromosome 1 and 5: their syntenic line almost kept stable when MILE value was changed.

It is obvious that large-scale duplication blocks were detected in both alignments when the 30 aa MILE value was chosen (Fig.2a and Fig.2b). According to the reference table (Table 1), 30 aa MILE corresponds to divergent time of 84 million years ago, i.e. many (~50%) genes appeared 84 million years ago and most genes (>95%) that appeared 105 million years ago, were ignored or scanned off, and only the genes originated within 84 millions years were focused on. The results showed that the two large-scale duplicate events both occurred 84 million years ago. But at 100 aa MILE level, only one duplicate block between chromosome 11 and 12 was detected (Fig.2d), which indicated that the duplicate event occurred more recently, about 16 million years ago, and another duplicate event occurred even earlier, around 21−84 million years ago. The results also implied that two large-scale duplication events should occur during early evolution of rice genome. All chromosomes of rice were further detected using PROmer with 30 and 100 MALE, respectively (See complementary materials at http://ibi.zju.edu.cn/bioinplant/data/). Many duplicate blocks were detected at 30 MALE levels, as in chromosome 1 and 5. The blocks did not overlap each other and covered most parts of rice chromosomes. The results suggested that beside the recent duplication (~5 Mb) of chromosome 11 and 12, a whole-genome duplication occurred in rice genome about 20−84 million years ago. Our results were consistent with those of Paterson *et al.*(2004), who suggested that a whole-genome duplication and a duplicate event between chromosome 11 and 12 occurred ~70 million years ago and recently, respectively.

DISCUSSION

Our studies on MALE changes of main gene families from four representative species during evolution indicated that MALEs are related to divergent time significantly during early evolution of gene families, and a reference table between MALE and its corresponding divergent time was set up. An important application of the table is large-scale genome analysis, such as genome alignment. A successful example of application of the table to detect the large-scale duplication events in rice genome is given in this paper. The release of the high-quality genome sequence of various organisms makes possible the analysis of chromosomal behavior and other evolutionary processes. For large-scale genome sequence analysis, our reference table between MALE and divergent time provides an estimate of divergent times of target genes concerned. But it must be cautioned that the table is associated with a large degree of uncertainty when it deals with ancient (d_A>0.5) divergent events of duplicated genes because the MALEs of those ancient duplicate genes were not sensitive to divergent time (Figs.1a−1c).

A rational default value of MILE for genome alignment should be set to guarantee that hits with biological significance are returned. Our analysis on random protein sequences indicated that few (<1.1%) MALE between random sequences were over 6 aa. Therefore, 6 aa seems to be a good starting point for genome alignment. Of course, more strictly speaking, length (such as 7 aa, less than 0.04% of total random MALEs are over this length) could also be chosen for some particular analysis.

Another traditional evolutionary scale is nucleotide acid synonymous substitutions rate (*K*s), which is widely used in analysis of evolutionary divergence of genes. In this study, the scale was not used because higher *K*s values (*K*s<2.0, even 1.0) are associated with a large degree of uncertainty due to saturation of substitutions (Li, 1997; Blanc and Wolfe, 2004).

ACKNOWLEDGMENTS

We thank Prof. Mingwei Gao and Dr. Shaukat Ali for their critical reading.

References

Arratia, R., Gordon, L., Waterman, M.S., 1986. An extreme value theory for sequence matching. *Ann. Stat.*, **14**:971-993.

Blanc, G., Wolfe, K.H., 2004. Widespread paleopolyploidy in model plant species inferred from age distributions of duplicate genes. *Plant Cell*, **16**:1667-1678.

Brown, T.A., 1999. Genomes. BIOS Scientific Publishers Limited, Oxford.

Delcher, A.L., Phillippy, A., Carlton, J., Salzberg, S.L., 2002. Fast algorithms for large-scale genome alignment and comparison. *Nucleic Acids Res*., **30**:2478-2483.

Goff, S.A., Ricke, D., Lan, T.H., Presting, G., Wang, R., Dunn, M., Glazebrook, J., Sessions, A., Oeller, P., Varma, H., *et al.*, 2002. A draft sequence of the rice genome (*Oryza sativa* L. ssp. *japonica*). *Science*, **296**:92-100.

Karlin, S., Altschul, S.F., 1990. Methods for assessing the statistical significance of molecular sequence features by using general scoring schemes. *Proc. Natl. Acad. Sci. USA*, **87**(6):2264-2268.

Li, W.H., 1997. Molecular Evolution. Sinauer Associates, Sunderland, MA.

Lynch, M., Conery, J.S., 2000. The evolutionary fate and consequences of duplicate genes. *Science*, **290**: 1151-1155.

Ohno, S., 1970. Evolution by Gene Duplication. George Allen and Unwin, London.

Paterson, A.H., Bowers, J.E., Chapman, B.A., 2004. Ancient polyploidization predating divergence of the cereals, and its consequences for comparative genomics. *Proc. Natl. Acad. Sci. USA*, **101**:9903-9908.

Yang, Z., 1999. Phylogenetic Analysis by Maximum Likelihood (PAML). Version 2., London, UK.