Nucleic Acids Research, 2007, Vol. 35, Web Server issue ***W121–W125***
doi:10.1093/nar/gkm426

FGF: A web tool for Fishing Gene Family in a whole genome database

Hongkun Zheng[1,2], Junjie Shi[1], Xiaodong Fang[1], Yuan Li[1], Søren Vang[3], Wei Fan[1], Junyi Wang[4], Zhang Zhang[1], Wen Wang[5], Karsten Kristiansen[2] and Jun Wang[1,2,6,*]

[1]Beijing Institute of Genomics of Chinese Academy of Sciences, Beijing Genomics Institute, Beijing 101300, China, [2]Department of Biochemistry and Molecular Biology, University of Southern Denmark, DK-5230, Odense M, Denmark, [3]Research Unit for Molecular Medicine, Aarhus University Hospital and Faculty of Health Sciences, University of Aarhus, 8200 Aarhus N, Denmark, [4]The State Key Laboratory of Plant Cell and Chromosome Engineering, Institute of Genetics and Developmental Biology, Chinese Academy of Sciences, Beijing 100101, China, [5]CAS-Max-Plank Junior Research Group, Key Laboratory of Cellular and Molecular Evolution, Kunming Institute, China and [6]The Institute of Human Genetics, University of Aarhus, DK-8000 Aarhus C, Denmark

Received July 15, 2006; Revised May 3, 2007; Accepted May 9, 2007

ABSTRACT

Gene duplication is an important process in evolution. The availability of genome sequences of a number of organisms has made it possible to conduct comprehensive searches for duplicated genes enabling informative studies of their evolution. We have established the FGF (Fishing Gene Family) program to efficiently search for and identify gene families. The FGF output displays the results as visual phylogenetic trees including information on gene structure, chromosome position, duplication fate and selective pressure. It is particularly useful to identify pseudogenes and detect changes in gene structure. FGF is freely available on a web server at http://fgf.genomics.org.cn/

INTRODUCTION

More than 35 years ago, Susumu Ohno stated that gene duplication was the single most important factor underlying genome evolution (1). Much work has been performed delineating the relationship between gene duplication and organism evolution (2–4), and the recent rise in whole-genome sequence data makes it possible to characterize entire sets of duplicated genes within one species (5). Sequence alignment software such as 'BLAST' (6) can be used to find homologous genes, and from these, orthologous and paralogous genes can be identified by phylogenetic analyses. Such tools facilitate studies on gene duplications (7–9), production of new genes (10–12) and distribution of pseudogenes (13–15). By estimating the ages of a large set of duplicates it is possible to estimate the birth and death rates for individual genes.

In order to study the relation between gene duplication and evolution, we have made a web server-based tool for identification and analysis of gene duplications in selected genomes. The FGF server uses protein sequences as bait for fishing gene families and extracting related information, and thus we have named the tool Fishing Gene Family (FGF). The FGF program visualizes the chromosomal position of the duplications, the exon–intron structure and constructs a phylogenetic tree based on a distance matrix. By analyzing stop codons, frame shift truncations and the ratio of non-synonymous to synonymous nucleotide substitution rates (16–18) (*Ka*/*Ks* ratio) of every copy, the functional outcome of the duplication is predicted. The FGF tool has been implemented and validated by an analysis of 13 089 proteins from the rice strain *Indica* 93-11 (12).

IMPLEMENTATION

The FGF server is implemented in JSP + MySQL. The web interface is displayed in Figure 1. Upon submission of a task, the user must provide a valid e-mail address or register and login on the server. Detailed information on the required format of input sequences, parameter settings and possibilities for adjusting parameters are available online by addressing the user's guide or by clicking on relevant question marks. After completion of the task, the user receives an e-mail with a job ID for accessing and downloading the results.

The computation process of the FGF server is shown in Figure 2. Initially, the program searches for copies of the

*To whom correspondence should be addressed. Tel: +86 10 8048 1664; Fax: +86 10 8049 8676; Email: wangj@genomics.org.cn
Correspondence may also be addressed to Karsten Kristiansen. Tel: +45 6550 2408; Fax: +45 6550 2467; Email: kak@bmb.sdu.dk

The authors wish it to be known that, in their opinion, the first three authors should be regarded as joint First Authors

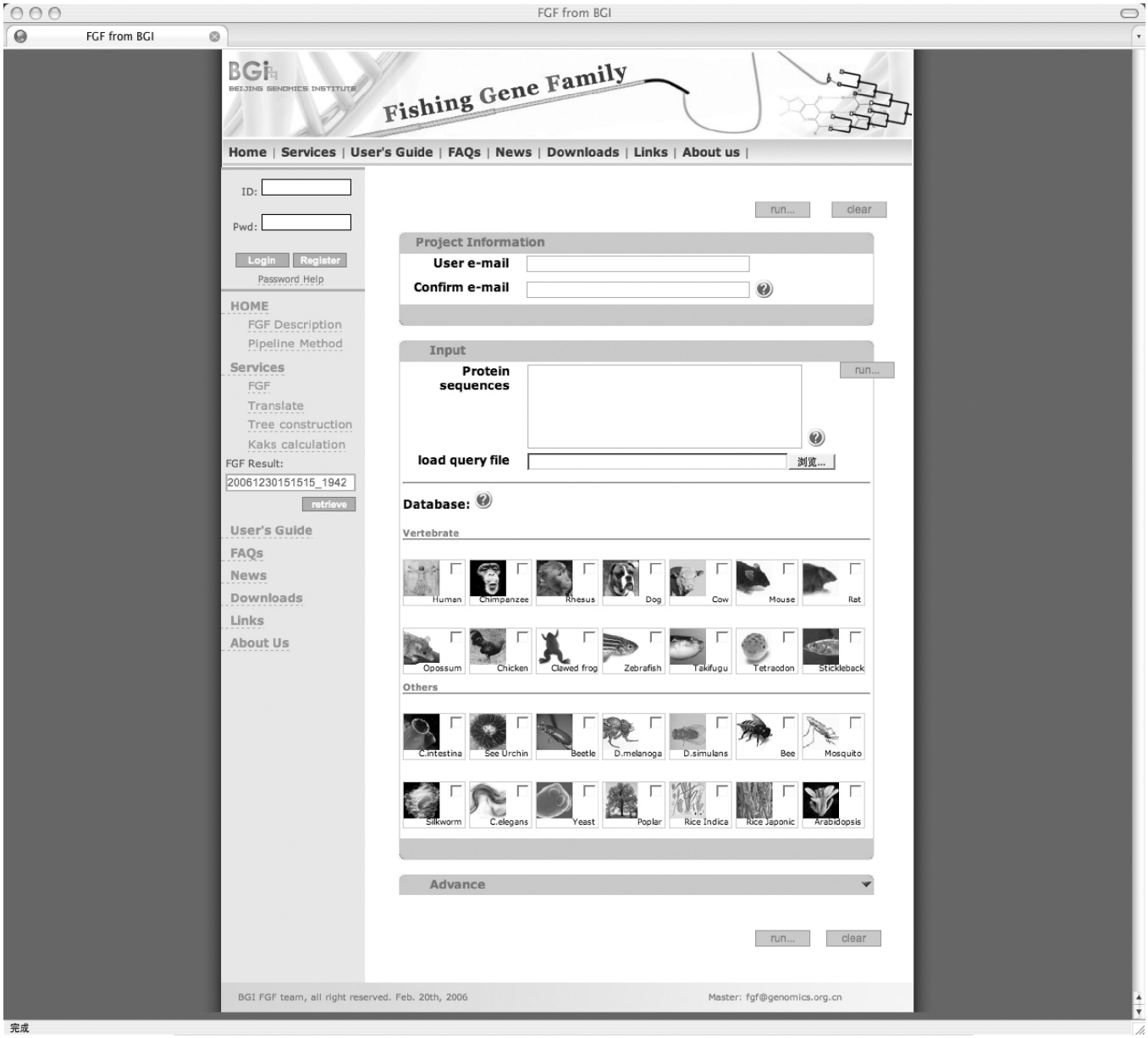

Figure 1. The FGF server web page. More details are available in the user's guide.

query protein in a genome using the tBLASTn program in the BLAST package, and then joins sequence blocks using a dynamic programming algorithm as follows: Each BLAST block has a score which reflects its length and identity; longer length and higher identity will have a higher score. If the BLAST blocks have <20% overlap, they will be joined. There is a score-penalty for the gaps between the BLAST blocks, a longer gap has a higher score-penalty. Following this procedure, we obtain BLAST block chains having the highest score (for more details see the website help). The program next realigns the query protein to homologous regions in the genome using 'GeneWise' (19). After filtering out false paralogs, basic information on the duplicated genes such as sequence, structure, position and premature stop codons/frame shifts are distilled from the 'GeneWise' results. The gene family is subsequently presented as a phylogenetic tree using 'njtree', the core engine of 'TreeFam' (20), in which duplication basic information and evolutionary information such as selective pressure (*Ka*/*Ks* ratio) are included. To calculate *Ka*/*Ks* ratios we first calculate pairwise *Ka* and *Ks* distances between each pair of sequences, which yields two distance matrices. Then we fix the topology of the phylogenetic tree and estimate branch lengths with the constrained neighbor-joining method. This algorithm was designed in the development of TreeFam database where automatic trees must agree with curated trees. The *Ka*/*Ks* ratio is then calculated from the *Ka* and *Ks* branch distances. We provide default parameters for alignment and evolutionary analysis, which can be altered freely by the users.

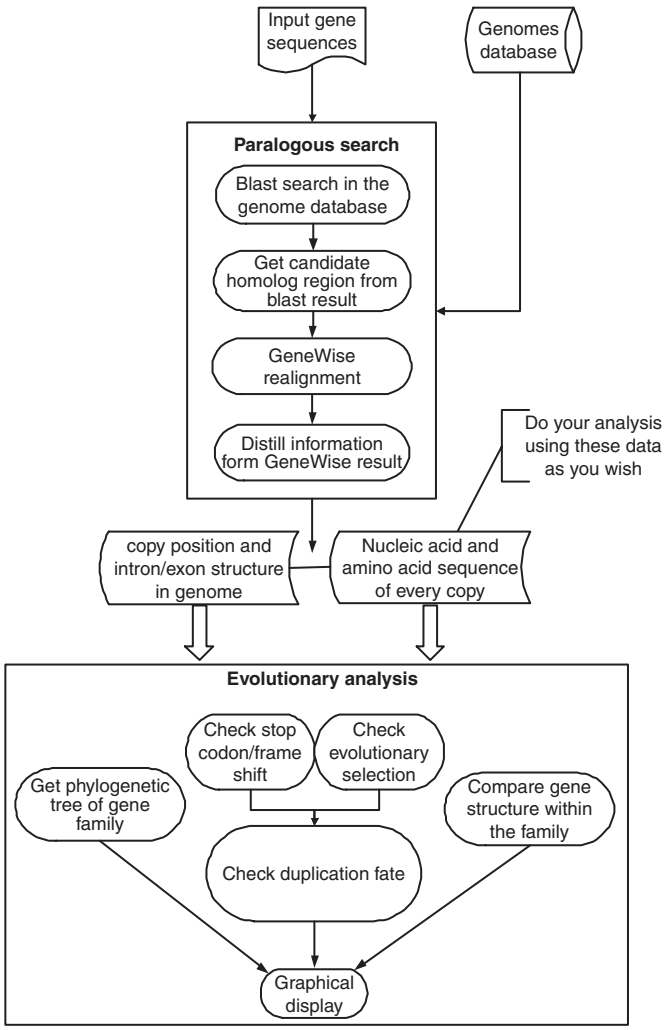

Figure 2. Flowchart of Fishing Gene Family. The flowchart mainly involves two parts. The fist step is paralogous search defining the gene family using BLAST candidate homologs from genome searches followed by accurate alignments using 'GeneWise'. After filtering out false paralogs, basic information on the duplicated genes such as sequence, structure, position and premature stop codons/frame shifts are distilled from the 'GeneWise' results. The gene family is subsequently presented as a phylogenetic tree using 'njtree', the core engine of 'TreeFam', in which duplication basic information and evolutionary information such as selective pressure (*Ka*/*Ks* ratio) are included. Default parameters for alignment and evolutionary analysis can freely be altered by the user.

To ensure the highest validity in the gene family finding, FGF uses the following four restrictions: (i) A 'BLAST' E-value is used to tune the similarity threshold below which gene pairs are eliminated; (ii) the maximum gap length and cutoff setup can filter out short and false sequences leaving only sequences with sufficient similarity to the protein; (iii) a second alignment software, 'GeneWise', filters out further unqualified sequences and (iv) when different gene duplications share the same domain and the overlap is >60% of length, only the best homolog with the highest identity is kept according to the relevant parameters selected by the user. The 'TreeFam' method is used to build phylogenetic trees and give the bootstrap of each node as a reflection of the stability of the tree

USAGE

Case studies

To illustrate the use of the FGF server, examples based on protein data from human and rice are shown. We have analyzed the gene family of the Hsp90 protein (GI: 40254816) (21) in human and AK061894 (22) in rice. The resulting FGF graphics are shown in Figure 3. A red star marks the best hit to the query protein, and identity of the duplications of the query protein is color-coded. Stop codon/frame-shifts, *Ka*/*Ks* ratios and bootstrap values are all denoted in the phylogenetic tree. The distance based on *Ka* or *Ks*, and the sequence lengths are marked as abscissa in the figure. In sum, Figure 3 provides the following information: (i) the copy number of the gene family: 12 members in the rice AK061894 protein family and 15 members in the human Hsp90 protein family; (ii) the fate of the gene duplications estimated by stop codon/frame shift truncations and *Ka*/*Ks* ratio; (iii) the changes of gene structure comparing exon numbers and intron length and (iv) a phylogenetic tree reflecting the evolutionary relationship of member of the gene family.

The fate of the duplication. FGF provides information on stop codons and frame shifts truncations in duplicated genes from which the fate of the duplications can be assessed. The presence of a stop codon in one sequence AK061894_Chr02_652556_656933 and frame shifts in both duplications AK061894_Chr10_3084964_3089556 and AK061894_Chr02_652556_656933 in the rice AK061894 gene family indicate that these duplications are pseudogenes (Figure 3a). In the human Hsp90 protein family, 12 duplications have frame shifts and stop codons and the numbers of frame shifts/stop codons are ≥4 indicating that these duplications are probably non-functional (Figure 3b). An analysis of gene retroposition duplication in the human genome has indicated that ~85% of retroposed gene duplications represent pseudogenes (23). In contrast, our previous research using FGF indicated that only 25% of retroposed gene duplications in the rice genome represent truncation mutations (12), suggesting that most duplications probably are functional. By inspecting the *Ka*/*Ks* ratio, especially in the human example, we observed six members with *Ka*/*Ks* ratios higher than 0.8, and four for which no ratio can be calculated because of the high number of mutations. Figure 3 illustrates the marked differences between the duplication fate of rice and human genes; 12 out of 15 duplications in human, but only 2 out of 11 duplications in rice have lost their function.

The change of gene structure. Often the structure of a gene is changed after duplication. By using the FGF software, the different changes can easily be visualized. Figure 3 shows that the AK061894 and the Hsp90 gene families are composed of both multi-exon and single-exon members, and that the exon numbers are different among the family members. In the AK061894 gene family, the exon number of AK061894_Chr02_1209004_1213519, AK061894_Chr10_3084964_3089556, AK061894_Chr05_171983_177329 and AK061894_Chr06_31782427_31787064

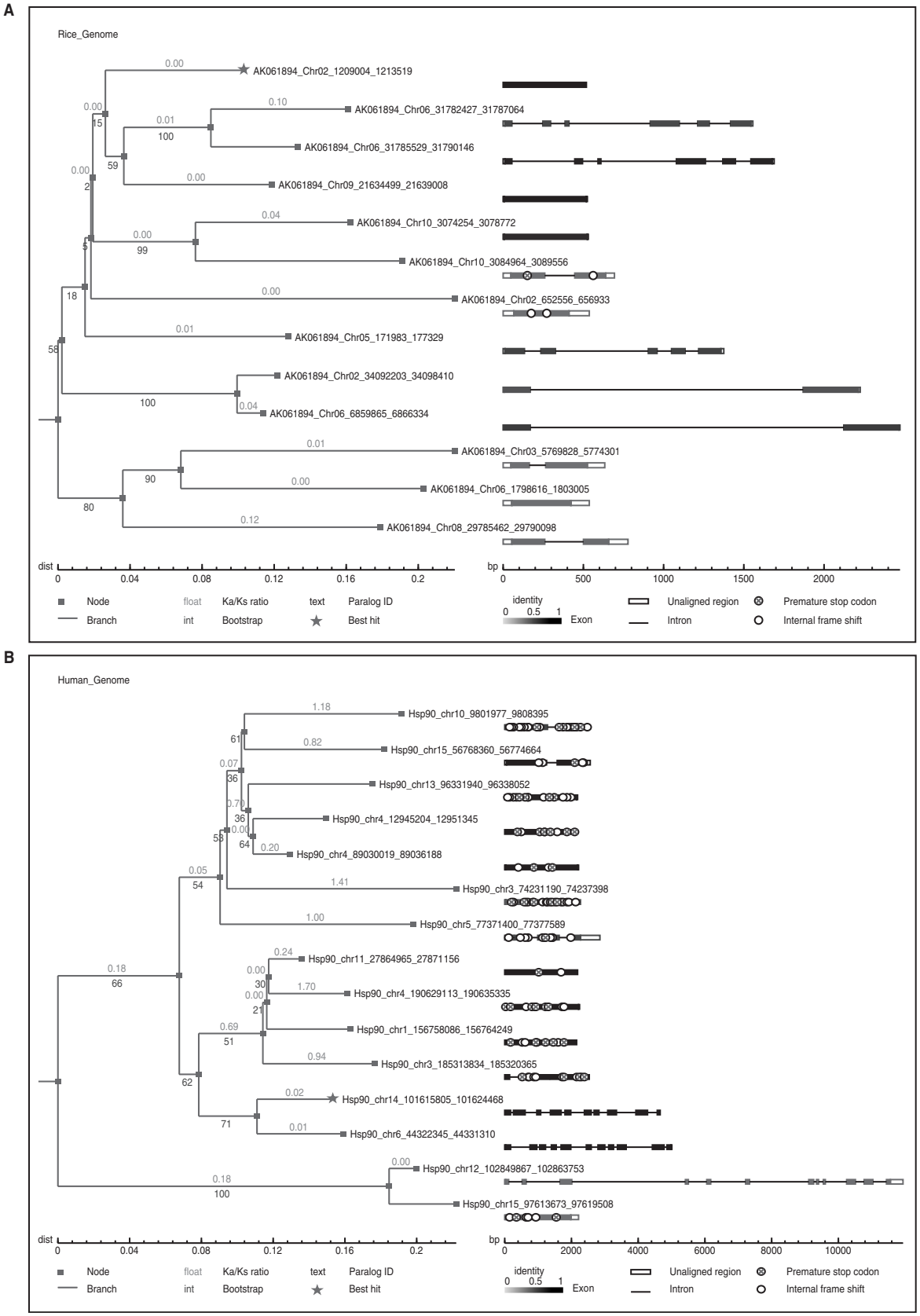

Figure 3. Two examples of FGF graphic result. (**A**) Evolutionary information of gene AK061894 (GI:32971912, potential cds 95–613) in rice genome and (**B**) Evolutionary information of gene Hsp90 (GI: 40254816) in human genome.

are 1, 2, 5 and 6, respectively, and the gene structures are changed dramatically. Four duplications are single exon genes. These are possibly generated by retroposed duplications since there are other genes in the family with multiple exons.

In some duplications, the exon number is the same, but the intron size has changed. In the rice AK061894 gene family, AK061894_Chr06_31782427_31787064 and AK061894_Chr06_31785529_31790146 both contain six exons but the first intron in each differs in lengths. AK061894_Chr02_34092203_34098410 and AK061894_Chr06_6859865_6866334 both have two exons but their intron sizes have changed. These two gene pairs are closely connected in the phylogenetic tree. The same phenomenon was found in the human Hsp90 protein family.

CONCLUSION

The FGF web server integrates homologous search and evolutionary analysis and distills useful information to present the results in an easily accessible graphic representation. The tool deciphers not only the members of a gene family, but also their evolutionary relationships and helps in deducing the fate of the duplications. The software tries to restrict some inaccurate factors when searching for gene duplication, building a phylogenetic tree and calculating the *Ka*/*Ks* ratios to ensure the validity of the results. A 'TreeFam' method is used to calculate the phylogenic trees in order to achieve the most accurate phylogenic trees even for larger families. Because of its simplicity in the input, the FGF software is suitable for researchers studying evolution and bioinformatics.

ACKNOWLEDGEMENTS

This project was supported by the Chinese Academy of Sciences (KSCX2-YW-N-023; GJHZ0701-6), the Ministry of Science and Technology under high-tech program 863 (2006AA02Z334; 2006AA10A121), the Beijing Municipal Science and Technology Commission (D07030200740000) and the National Natural Science Foundation of China (90608010; 90208019; 90403130; 30221004; 90612019; 30392130). Other support came from Ole Rømer grants from the Danish Natural Science Research Council. We thank Richard Durbin from the Wellcome Trust Sanger Institute for thoughtful suggestions. We thank Heng Li for providing software, Jue Ruan and Yafeng Hu for helping with the web construction, Manyuan Long from the University of Chicago and Shunping He from the Institute of Hydrobiology of Chinese Academy of Sciences for helping with the test process. Funding to pay the Open Access publication charges for this article was provided by the Danish Medical Research Council.

Conflict of interest statement. None declared.

REFERENCES

1. Ohno,S. (1970) *Evolution by Gene Duplication*. Springer-Verlag New York, 150.
2. Ferris,S.D. and Whitt,G.S. (1979) Evolution of the differential regulation of duplicate genes after polyploidization. *J. Mol. Evol.*, **12**, 267–317.
3. Iwabe,N., Kuma,K. and Miyata,T. (1996) Evolution of gene families and relationship with organismal evolution: rapid divergence of tissue-specific genes in the early evolution of chordates. *Mol. Biol. Evol.*, **13**, 483–493.
4. Lundin,L.G. (1999) Gene duplications in early metazoan evolution. *Semin. Cell. Dev. Biol.*, **10**, 523–530.
5. Friedman,R. and Hughes,A.L. (2001) Gene duplication and the structure of eukaryotic genomes. *Genome Res.*, **11**, 373–381.
6. Altschul,S.F., Madden,T.L., Schaffer,A.A., Zhang,J., Zhang,Z., Miller,W. and Lipman,D.J. (1997) Gapped BLAST and PSI-BLAST: a new generation of protein database search programs. *Nucleic Acids Res.*, **25**, 3389–3402.
7. Wolfe,K.H. and Shields,D.C. (1997) Molecular evidence for an ancient duplication of the entire yeast genome. *Nature*, **387**, 708–713.
8. Lynch,M. and Conery,J.S. (2000) The evolutionary fate and consequences of duplicate genes. *Science*, **290**, 1151–1155.
9. Borrelli,L., De Stasio,R., Filosa,S., Parisi,E., Riggio,M., Scudiero,R. and Trinchella,F. (2006) Evolutionary fate of duplicate genes encoding aspartic proteinases. Nothepsin case study. *Gene*, **368**, 101–109.
10. Long,M., Betran,E., Thornton,K. and Wang,W. (2003) The origin of new genes: glimpses from the young and old. *Nat. Rev. Genet.*, **4**, 865–875.
11. Long,M. and Langley,C.H. (1993) Natural selection and the origin of jingwei, a chimeric processed functional gene in Drosophila. *Science*, **260**, 91–95.
12. Wang,W., Zheng,H., Fan,C., Li,J., Shi,J., Cai,Z., Zhang,G., Liu,D., Zhang,J. *et al.* (2006) High rate of chimeric gene origination by retroposition in plant genomes. *Plant Cell*, **18**, 1791–1802.
13. Harrison,P.M., Echols,N. and Gerstein,M.B. (2001) Digging for dead genes: an analysis of the characteristics of the pseudogene population in the Caenorhabditis elegans genome. *Nucleic Acids Res.*, **29**, 818–830.
14. Coin,L. and Durbin,R. (2004) Improved techniques for the identification of pseudogenes. *Bioinformatics*, **20**(Suppl. 1), I94–I100.
15. Torrents,D., Suyama,M., Zdobnov,E. and Bork,P. (2003) A genome-wide survey of human pseudogenes. *Genome Res.*, **13**, 2559–2567.
16. Nei,M. and Gojobori,T. (1986) Simple methods for estimating the numbers of synonymous and nonsynonymous nucleotide substitutions. *Mol. Biol. Evol.*, **3**, 418–426.
17. Yang,Z. (1997) PAML: a program package for phylogenetic analysis by maximum likelihood. *Comput. Appl. Biosci.*, **13**, 555–556.
18. Zhang,Z., Li,J. and Yu,J. (2006) Computing Ka and Ks with a consideration of unequal transitional substitutions. *BMC Evol. Biol.*, **6**, 44.
19. Birney,E., Clamp,M. and Durbin,R. (2004) GeneWise and Genomewise. *Genome Res.*, **14**, 988–995.
20. Li,H., Coghlan,A., Ruan,J., Coin,L.J., Heriche,J.K., Osmotherly,L., Li,R., Liu,T., Zhang,Z. *et al.* (2006) TreeFam: a curated database of phylogenetic trees of animal gene families. *Nucleic Acids Res.*, **34**, D572–D580.
21. Wesche,J., Malecki,J., Wiedlocha,A., Skjerpen,C.S., Claus,P. and Olsnes,S. (2006) FGF-1 and FGF-2 require the cytosolic chaperone Hsp90 for translocation into the cytosol and the cell nucleus. *J. Biol. Chem.*, **281**, 11405–11412.
22. Kikuchi,S., Satoh,K., Nagata,T., Kawagashira,N., Doi,K., Kishimoto,N., Yazaki,J., Ishikawa,M., Yamada,H. *et al.* (2003) Collection, mapping, and annotation of over 28,000 cDNA clones from japonica rice. *Science*, **301**, 376–379.
23. Zhang,Z., Harrison,P.M., Liu,Y. and Gerstein,M. (2003) Millions of years of evolution preserved: a comprehensive catalog of the processed pseudogenes in the human genome. *Genome Res.*, **13**, 2541–2558.