EPGD: a comprehensive web resource for integrating and displaying eukaryotic paralog/paralogon information

Guohui Ding¹,², Yan Sun¹,², Hong Li¹,², Zhen Wang¹,², Haiwei Fan¹, Chuan Wang¹, Dan Yang⁴ and Yixue Li¹,³,⁴*

¹Bioinformatics Center, Key Lab of Systems Biology, Shanghai Institutes for Biological Sciences, Chinese Academy of Sciences, 320 Yueyang Road, ²Graduate School of the Chinese Academy of Sciences, Shanghai 200031, ³Shanghai Center for Bioinformation Technology, 100 Qinzhou Road, Shanghai 200235 and ⁴Shanghai Information Center for Life Sciences, Shanghai Institutes for Biology Science, Chinese Academy of Science, Shanghai 200031, P. R. China

Received August 4, 2007; Revised October 8, 2007; Accepted October 10, 2007

ABSTRACT
Gene duplication is common in all three domains of life, especially in eukaryotic genomes. The duplicates provide new material for the action of evolutionary forces such as selection or genetic drift. Here we describe a sophisticated procedure to extract duplicated genes (paralogs) from 26 available eukaryotic genomes, to pre-calculate several evolutionary indexes (evolutionary rate, synonymous distance/clock, transition redundant exchange clock, etc.) based on the paralog family, and to identify block or segmental duplications (paralogons). We also constructed an internet-accessible Eukaryotic Paralog Group Database (EPGD; http://epgd.biosino.org/EPGD/). The database is gene-centered and organized by paralog family. It focuses on paralogs and evolutionary duplication events. The paralog families and paralogons can be searched by text or sequence, and are downloadable from the website as plain text files. The database will be very useful for both experimentalists and bioinformaticians interested in the study of duplication events or paralog families.

INTRODUCTION
The occurrences and consequences of gene and genome duplication events have been discussed for a long time (1,2). The duplication of genes and large genome regions (or entire genomes) is proposed to be an important mechanism for the evolution of phenotypic complexity, diversity and innovation, and as an origin of novel gene functions. To uncover the evolutionary trajectories of duplicated genes, previous studies have integrated transcriptomic, interactomic and other data (1). Such integrated approaches, focusing on gene duplications in genomes, have already contributed robust insights into important evolutionary questions, such as the complexity of genes (3), the evolution of genome architecture (4), growth of gene networks (5), the 2R hypothesis (6) and diversity of gene expression (7). Moreover, the duplicated genes can be used to investigate diverging gene functions, which, when allied with computational methods, may provide useful information for experimental approaches. An example is the analysis of the molecular basis of the adaptive evolution of the duplicated pancreatic ribonuclease gene in leaf-eating monkeys with both computational and experimental approaches (8).

As more genomes are examined, increasing evidences support the dominating role of gene duplication events in the expanding of genome content (2,9). A crucial step in the study of gene duplications is to identify duplicated genes (known as paralogs) in genome sequences and to distinguish these from genes that have similar sequence but arisen from convergent evolution or other mechanisms. Algorithm-based homology detection from primary sequences is the preferred approach to detect paralogs or paralogous regions (4).

In contrast to ortholog databases, there are only a few specific paralog databases available in the public domain. Even though several general homolog databases, such as Inparanoid (10), Ensembl Compara (11), NCBI

*To whom correspondence should be addressed. Tel: +86 21 54920089; Fax: +86 21 54920143; Email: yxli@sibs.ac.cn

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

© 2007 The Author(s)
This is an Open Access article distributed under the terms of the Creative Commons Attribution Non-Commercial License (http://creativecommons.org/licenses/by-nc/2.0/uk/) which permits unrestricted non-commercial use, distribution, and reproduction in any medium, provided the original work is properly cited.
homologene (12), include some paralog information, they
did not comprehensively summarize and display the
evolution information of paralogs. In order to construct
a stable web resource that supports easy browsing and
downloading of evolutionary information on paralogous
genes, we created EPGD (Eukaryotic Paralog Group
Database; http://epgd.biosino.org/EPGD/). Several steps
used to identify the paralogs contained in the EPGD were
used previously to detect the duplication events in the
family of animal transmembrane genes (13). Using this
work (13) as a basis, we developed a semi-automatic
procedure for collecting the within-species paralog
families from genomes and pre-calculating several evolu-
tionary indexes of these families. We collected the
paralogs only from eukaryotes, as they are known to
have a higher rate of gene duplication than Prokaryotes
(14) and are more widely studied in this field.

A pioneer in the construction of paralog database is
paraDB (15). A highlight of paraDB is the display of
paralogons, which have been thoroughly investigated in the
human genome (16) and are reviewed by Van de Peer (4).
EPGD inherits this feature and adopts the term 'para-
logon', defined as homologous genomic segments created
by partial or complete genome duplication. EPGD focuses
on families of paralogs and integrates spatial and temporal
data to diagnose gene duplication processes comprehen-
sively (17). The ratio of dN (the rate of non-synonymous
substitutions) to dS (the rate of synonymous substitutions)
(18), synonymous distance/clock, transition redundant
exchange (TREx) clock (19), paralogons and several
other features were generated by computational methods
and deposited in the database.

In the current EPGD version, 26 eukaryotic genomes
were processed and 35991 paralog families and 29480
paralogons were identified and stored (Table 1). To our
knowledge, it is one of the most extensive paralog
databases in public domain. All data can be browsed,
searched and downloaded directly from the website.

### CONSTRUCTION AND CONTENT

EPGD is implemented through MySQL relational
database (http://www.mysql.com) and JavaServer Pages
technology (http://java.sun.com/products/jsp/). The raw
datasets of 26 eukaryotic genomes (Table 1) in GeneBank
flat file format (GBK) were downloaded from the NCBI
FTP site (ftp://ftp.ncbi.nih.gov/genomes) in March 2007.
Proteins, coding sequences (CDS) and gene location
information were extracted from these GBK files with a
PERL script.

#### Overview of the procedure

A total of 531 715 coding sequences and corresponding
proteins were obtained after preprocessing. Only the
protein sequences were used to construct the paralog
families. The procedure is briefly described below:

(i) Pairwise alignments of the proteins using gapped
BLAST (20), with filtering for low sequence
complexity regions using SEG (21). The default
parameters were used, except for the threshold
E-value of $10^{-5}$.

| Species                  | TaxID   | Paralog | Gene     | Paralogon | Ratioa | Family | Family size\(^{b}\) |
|--------------------------|---------|---------|----------|-----------|--------|--------|-------------------|
| Plasmodium falciparum    | 36 329  | 494     | 5365     | 433       | 0.09   | 90     | 5.4889            |
| Khayveromyces Lactis     | 284 590 | 539     | 5504     | 50        | 0.1    | 206    | 2.6165            |
| Cryptococcus neoformans  | 214 684 | 736     | 6617     | 94        | 0.11   | 252    | 2.9206            |
| Apis mellifera           | 7460    | 1223    | 9430     | 58        | 0.13   | 371    | 3.2965            |
| Dekaryomyces Hansenii    | 284 592 | 992     | 7081     | 109       | 0.14   | 334    | 2.9701            |
| Candida glabrata         | 284 593 | 756     | 5534     | 72        | 0.14   | 304    | 2.4868            |
| Yarrowia lipolytica      | 284 591 | 1056    | 7180     | 317       | 0.15   | 294    | 3.5918            |
| Schizosaccharomyces pombe| 284 812 | 815     | 5374     | 119       | 0.15   | 302    | 2.6987            |
| Encephalitozoon cuniculi | 284 813 | 312     | 2029     | 161       | 0.15   | 87     | 3.5862            |
| Aspergillus fumigatus    | 530 879 | 1573    | 10 157   | 470       | 0.15   | 504    | 3.121             |
| Anopheles gambiae        | 180 454 | 2169    | 13 748   | 521       | 0.16   | 565    | 3.8389            |
| Bos taurus               | 9913    | 4995    | 28 806   | 541       | 0.17   | 1232   | 4.0544            |
| Danio rerio              | 7955    | 6765    | 3863     | 1014      | 0.18   | 1915   | 3.5326            |
| Saccharomyces cerevisiae | 4932    | 1269    | 6198     | 484       | 0.2    | 473    | 2.6829            |
| Drosophila melanogaster  | 7227    | 3130    | 14 838   | 568       | 0.21   | 773    | 4.0492            |
| Macaca mulatta           | 9544    | 6579    | 29 122   | 1189      | 0.23   | 1826   | 3.603             |
| Pan troglodytes          | 9398    | 7147    | 31 482   | 1913      | 0.23   | 1944   | 3.6764            |
| Tribolium castaneum      | 7070    | 2335    | 9837     | 344       | 0.24   | 549    | 4.2532            |
| Gallus gallus            | 9031    | 5017    | 19 828   | 883       | 0.25   | 1500   | 3.3447            |
| Canis familiaris         | 9615    | 6065    | 20 053   | 1443      | 0.3    | 1671   | 3.6296            |
| Caenorhabditis elegans   | 6239    | 6528    | 21 052   | 1139      | 0.31   | 1331   | 4.9046            |
| Homo sapiens             | 9606    | 10 962  | 33 610   | 2134      | 0.33   | 3445   | 3.182             |
| Mus musculus             | 10 090  | 14 592  | 41 323   | 2705      | 0.35   | 3390   | 4.3044            |
| Rattus norvegicus        | 10 116  | 12 959  | 35 786   | 2234      | 0.36   | 3387   | 3.8261            |
| Arabidopsis thaliana     | 3702    | 15 573  | 32 025   | 9581      | 0.49   | 3590   | 4.3379            |
| Strongylocentrotus purpuratus | 7668  | 15 773  | 30 552   | 904       | 0.52   | 5656   | 2.7887            |

\(^{a}\)Ratio of the duplicated genes to all genes.
\(^{b}\)Average family size in genes.
The web interface was implemented using Java and Web interface.

Table 1.

Core tables. A summary of the data in EPGD is shown in indexes and annotation information, surround these two paralogons. The peripheral tables, i.e. evolutionary EPGD are the table of paralog families and the table of paralogous. Large datasets were obtained when the procedure was content in the database.

(ii) Definition of the homologous genes. Four criteria must be satisfied. (a) all high-scoring segment pairs (HSPs) in the target sequence have to be arranged in the same order as in the query protein sequence (22); (b) the remaining HSPs cover more than 80% of the protein length; (c) the similarity of each HSP is more than 50% (two amino acids are considered similar if their BLOSUM62 similarity score is positive) (22) and (d) these conditions are symmetrical for both genes.

(iii) Single linkage clustering of homologous genes (13). Generation of the primary paralog families.

(iv) Mapping the proteins to gene loci. Paralog families with at least two gene loci were retained.

(v) Multiple alignment of the proteins in each retained family. Clustalw (version 1.83) (23) was applied in this step.

(vi) Codon-level multiple alignment with the CDS in each family by using RevTrans (version 1.4) (24).

(vii) Calculations of the evolutionary indexes. \( d_N \) and \( d_S \) were calculated with the Nei and Gojobori (25) and the Yang and Nielsen methods (26), which were carried out using ydn00 from the PAML (Phylogenetic analysis by maximum likelihood) packages (27). The TREx distances were computed based on the definition (19): the fractional identity of silent sites in conserved 2-fold redundant codon sites, which was implemented by ourselves.

(viii) Construction of the arithmetic average (UPGMA) trees for grouping the proteins in a paralog family. These trees were derived from the \( d_S \) matrix, because the synonymous substitutions are thought to be approximative neutral molecular markers.

(ix) Identification of the paralogons using the algorithm developed by McLysaght et al. (16). Paralogons are two genomic segments that share a set of paralogous genes (4,16). After tandem duplications were masked, a greedy search algorithm was used to identify all paralogons between all pairs of chromosomes, based only on gene content but not gene order (4). Two criteria must be satisfied for a pair of paralogons. (a) they should contain at least two pairs of paralogous genes; (b) the gap size between two neighboring paralogous points in either chromosome should be less than the average length of 30 genes (16).

Content in the database

Large datasets were obtained when the procedure was applied to 26 genomes. We housed the data in a MySQL relational database. The kernel tables in the schema of EPGD are the table of paralog families and the table of paralogons. The peripheral tables, i.e. evolutionary indexes and annotation information, surround these two core tables. A summary of the data in EPGD is shown in Table 1.

Web interface

The web interface was implemented using Java and JavaServer Pages technologies. The user can inspect the datasets in the EPGD and see a summary of the current version. The records of paralog families, paralogons and genes (Figure 1) are randomly selected each time when ‘Glance’ page is visited (http://epgd.biosino.org/EPGD/glance.jsp).

As shown in Figure 1, if the gene record is obtained, the corresponding paralog family and paralogons can be linked from this page. The main content of the gene page (Figure 1A) starts with basic information of this gene (NCBI gene ID, taxonomy, EPGD family ID, location in the chromosome and simple description), followed by EPGD paralogons, which include or cover this gene. We defined that a gene is ‘included’ in a paralogon if it has at least one corresponding paralog in this paralogon region (paralogon-defining gene), while a gene is ‘covered’ by a paralogon if it does not have any corresponding paralog in this paralogon region (paralog-intervening gene). The coding sequences of the gene are listed at the bottom of the page.

The outline of the family page is similar to that of gene page (Figure 1B). Multi-aligned sequences in protein or codon level, pre-calculated evolution indexes \( [d_N, d_S, d_{S}, d_{E}, d_{F}, d_{Z}, d_{E}, d_{G}, d_{H}, d_{I}, d_{J}, d_{K}, d_{L}, d_{M}, d_{N}, d_{O}, d_{P}, d_{Q}, d_{R}, d_{S}, d_{T}, d_{U}, d_{V}, d_{W}, d_{X}, d_{Y}, d_{Z}] \) and UPGMA tree based on \( d_S \) were displayed on this page. The multi-alignments can be viewed in plain text or be displayed with the Jalview alignment viewer (28) (Figure 1). In the page which is hyperlinked from ‘Evolution indexes of Pairwise CDSs’, a row with a \( d_N/d_S \) different from the neutral expectation of 1 (\( z > 1.96 \) or \( z < -1.96 \)) is color coded orange (Figure 1). The \( z \) score is computed using equation (18)

\[
z = \frac{d_N - d_S}{\sqrt{SE_{d_N}^2 + SE_{d_S}^2 - Cov(d_N,d_S)}}
\]

where \( z \) is the \( z \) score, \( d_N \) is the rate of non-synonymous substitutions, \( d_S \) is the rate of synonymous substitutions, \( SE_{d_N} \) and \( SE_{d_S} \) are the standard errors of \( d_N \) and \( d_S \), and \( Cov(d_N,d_S) \) is the covariance of \( d_N \) and \( d_S \). We assume that the non-synonymous substitutions and the synonymous substitutions are independent and set \( Cov(d_N,d_S) \) to zero (18).

The main part of the paralogon page contains basic information (taxonomy, locations in the chromosomes, average block length, average block density, number of links) of the paralogon, followed by an image thumbnail displaying a graphic view of the paralogon. Here, ‘the average block density’ is the arithmetic mean of the ratio of paralogon-defining genes to all genes in both sides of the paralogon; ‘number of links’ is the number of unique paralog families linked in the paralogon region. When the mouse hovers over this thumbnail, an enlarged view of this image pops up. Gene names and their regions in the enlarged graphic view of this paralogon are hyperlinked to the gene records in database.

The user can access the records in the EPGD with customized queries (Figure 2). From the ‘iSearch’ webpage (Figure 2A), ‘any text’ and nucleic acid or protein sequences can be searched without setting any parameter. Advanced Search pages with numerous input options (Figure 2B and C) can be accessed via the links (‘Advanced Text Search’ or ‘Advanced Sequence Search’).
from 'iSearch' page. The sequence search is powered by NCBI Blast package (20). Each search returns a result list of records in the database, which provides the hyperlinks to detailed pages (Figure 2D).

DATA AVAILABILITY
The EPGD is available for download through the 'DOWNLOADS' link in the website as a FASTA file containing all proteins, family members lists, evolutionary indexes and paralogon regions in plain text files.

RESULTS AND DISCUSSION
The properties of the paralog family spaces in EPGD
Table 1 gives a summary of the content of the current EPGD version. The proportions of duplicated genes in eukaryotes collected by EPGD range from 9%
(Plasmodium falciparum) to 52% (Strongylocentrotus purpuratus), and are smaller than previously reported (e.g. Homo sapiens, 38%; Arabidopsis thaliana, 65%; Drosophila melanogaster, 41%; Caenorhabditis elegans, 49%; Saccharomyces cerevisiae, 30%) (2). This is due to the rigorous criteria for paralog definition used to construct the EPGD and because many duplicated genes have eliminated characteristic signatures from their sequences during their evolution history (2). Since evolutionary indexes are highly unreliable for ancient gene duplications, rigorous criteria are essential for our database.

The size of the paralog families tends to be smaller than five genes. The distributions of paralog family size in all species of EPGD follow power law (data not shown) (29,30). As an example, Figure 4A displays the distribution of paralog family sizes in H. sapiens and the corresponding log–log diagram. The power law distribution indicates the robustness of our family detection method and the quality of gene prediction in the original data (29).

Consistent with previous studies on Bacteria and a small set of Eukarya (9,29,31), large genomes possess more paralog families and a higher proportion of genes belonging to paralog families than small genomes (Figure 3A and C). We find, however, only a weak correlation between the average size of families and the genome sizes (Figure 3B, \( r = 0.26, P = 0.19 \)), in contrast to the finding in Bacteria that average family size increases with genome size (31). This result suggests that the higher percentage of paralogs in large eukaryotic genome stems mainly from the emergence of new paralogon families. An expansion of existing gene families is not evident in Eukarya (Figure 3B).

The number of paralogons increases with the genome size (Figure 3D, \( r = 0.86, P = 3.356 \times 10^{-8} \)), indicating the effect of duplication of large genome segments on the evolution of genome size. Furthermore, the distribution of the paralogon size is also a skewed distribution (e.g. Figure 4D). Most of the paralogons have less than five linked families (98% of all human paralogons), because of the high level of gene loss after duplication, as well as recombination, chromosomal rearrangements and recombination. Still, the identification of putative paralogons provides many insights into evolutionary mechanisms (4).

The example of H. sapiens

Taking H. sapiens as an example (Figure 4), we plotted the distribution of paralog family size (Figure 4A), a scatter diagram of TREx distance versus dS (Figure 4B), a log–log graph of dN versus dS (Figure 4C) and the distribution of paralogon size (the number of linked families) (Figure 4D).
Transition redundant exchange (TREx) processes at the position of conserved 2-fold codon sites are thought to offer an approximation for a neutral molecular clock (19). We calculated the TREx distances for each paralog family, which provide a more homogeneous molecular clock than that provided by the dS. If the time since two genes diverged is long relative to the reciprocal of the rate constant with which these silent sites suffer transition substitutions, the TREx distance approximates 0.5. As seen from Figure 4B, TREx distances are negatively correlated with dS (Figure 4B, \( r = -0.89, P < 2.2 \times 10^{-16} \)). Therefore, the TREx distance can be used as an alternative of dS.

Similar to the work of Lynch et al. (32), dN was plotted as a function of dS (Figure 4C). The accumulation of non-neutral points when dS increases (Figure 4C) confirms the gradual increase of selective constraint on duplicates during evolutionary history (32). When dS is greater than 2, there are more points around the neutral expectation (Figure 4C). This is an artifact, resulting from the saturation effects in the estimation of dN and dS (33).

**PERSPECTIVES**

We plan to update EPGD every six months. As new eukaryotic organisms are fully sequenced and annotated, they will be added to EPGD using our procedure. In the future, ortholog annotation information will also be included. However, the development of the utilities for EPGD will still focus on tools for the analysis of duplication events, such as statistical tests of the paralogons (unpublished data) and chromosome ideograms. Furthermore, we will thoroughly analyze the data in EPGD and present insights into the effect of duplication events on genome evolution. The procedure to build the EPGD is currently semi-automatic. We will make the procedure totally automatic and start an open source project in the future.

**ACKNOWLEDGEMENTS**

We thank Zhonghao Yu, Xiaobin Xing, Yun Li, Kang Tu, Guangyong Zhen for helpful comments and suggestions. This research was supported by grants from National Basic Research Program of China (2006CB910700, 2004CB720103, 2004CB518606, 2003CB715901). Funding to pay the Open Access publication charges for this article was provided by National High-Tech R&D Program (863) (2006AA02Z334) and National Basic Research Program of China (2006CB910700, 2004CB720103, 2004CB518606, 2003CB715901).
Conflict of interest statement. None declared.

REFERENCES

1. Taylor, J.S. and Raes, J. (2004) Duplication and divergence: the evolution of new genes and old ideas. Annu. Rev. Genet., 38, 615–643.

2. Zhang, J. (2003) Evolution by gene duplication: an update. Trends Ecol. Evol., 18, 292–298.

3. He, X. and Zhang, J. (2005) Gene complexity and gene duplicability. Curr. Biol., 15, 1016–1021.

4. Van de Peer, Y. (2004) Computational approaches to unveiling ancient genome duplications. Nat. Rev., 5, 752–763.

5. Teichmann, S.A. and Babu, M.M. (2004) Gene regulatory network growth by duplication. Nat. Genet., 36, 492–496.

6. Makalowski, W. (2001) Are we polyploids? A brief history of one hypothesis. Genome Res., 11, 667–670.

7. Wagner, A. (2000) Decoupled evolution of coding region and mRNA expression patterns after gene duplication: implications for the neutralist-selectionist debate. Proc. Natl Acad. Sci. USA, 97, 6579–6584.

8. Zhang, J., Zhang, Y.P. and Rosenberg, H.F. (2002) Adaptive evolution of a duplicated pancreatic ribonuclease gene in a leaf-eating monkey. Nat. Genet., 30, 411–415.

9. Jordan, I.K., Makarova, K.S., Spouge, J.L., Wolf, Y.I. and Koonin, E.V. (2001) Lineage-specific gene expansions in bacterial and archaean genomes. Genome Res., 11, 555–565.

10. O’Brien, K.P., Remm, M. and Sonnhammer, E.L. (2005) Inparanoid: a comprehensive database of eukaryotic orthologs. Nucleic Acids Res., 33, D476–D480.

11. Hubbard, T.J., Aken, B.L., Beal, K., Ballester, B., Caccamo, M., Chen, Y., Clarke, L., Coates, G., Cunningham, F. et al. (2007) Ensembl 2007. Nucleic Acids Res., 35, D610–D617.

12. Wheeler, D.L., Barrett, T., Benson, D.A., Bryant, S.H., Canese, K., Chetvernin, V., Church, D.M., DiCuccio, M., Edgar, R. et al. (2007) Database resources of the National Center for Biotechnology Information. Nucleic Acids Res., 35, D5–D12.

13. Ding, G., Kang, J., Liu, Q., Shi, T., Pei, G. and Li, Y. (2006) Insights into the coupling of duplication events and macroevolution from an age profile of animal transmembrane gene families. PLoS Comput. Biol., 2, e102.

14. Lynch, M. and Conery, J.S. (2003) The origins of genome complexity. Science, 302, 1401–1404.

15. Leveau, M., Prat, K., Perrier, N., Birnbaum, D. and Coulier, F. (2003) ParaDB: a tool for paralogy mapping in vertebrate genomes. Nucleic Acids Res., 31, 63–67.

16. McLysaght, A., Hokamp, K. and Wolfe, K.H. (2002) Extensive genomic duplication during early chordate evolution. Nat. Genet., 31, 200–204.
17. Durand, D. and Hoberman, R. (2006) Diagnosing duplications – can it be done? Trends Genet., 22, 156–164.
18. Masatoshi Nei, S.K. (2000) Molecular Evolution and Phylogenetics. Oxford University Press, USA.
19. Benner, S.A. (2003) Interpretive proteomics – finding biological meaning in genome and proteome databases. Adv. Enzyme Regul., 43, 271–359.
20. 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.
21. Wootton, J.C. and Federhen, S. (1996) Analysis of compositionally biased regions in sequence databases. Methods Enzymol., 266, 554–571.
22. Perriere, G., Duret, L. and Gouy, M. (2000) HOBACGEN: database system for comparative genomics in bacteria. Genome Res., 10, 379–385.
23. Chenna, R., Sugawara, H., Koike, T., Lopez, R., Gibson, T.J., Higgins, D.G. and Thompson, J.D. (2003) Multiple sequence alignment with the Clustal series of programs. Nucleic Acids Res., 31, 3947–3950.
24. Wernersson, R. and Pedersen, A.G. (2003) RevTrans: multiple alignment of coding DNA from aligned amino acid sequences. Nucleic Acids Res., 31, 3537–3539.
25. Nei, M. and Gojobori, T. (1986) Simple methods for estimating the numbers of synonymous and nonsynonymous nucleotide substitutions. Mol. Biol. Evol., 3, 418–426.
26. Yang, Z. and Nielsen, R. (2000) Estimating synonymous and nonsynonymous substitution rates under realistic evolutionary models. Mol. Biol. Evol., 17, 32–43.
27. Yang, Z. (1997) PAML: a program package for phylogenetic analysis by maximum likelihood. Comput. Appl. Biosci., 13, 555–556.
28. Clamp, M., Cuff, J., Searle, S.M. and Barton, G.J. (2004) The Jalview Java alignment editor. Bioinformatics, 20, 426–427.
29. Enright, A.J., Kunin, V. and Ouzounis, C.A. (2003) Protein families and TRIBES in genome sequence space. Nucleic Acids Res., 31, 4632–4638.
30. Kunin, V., Teichmann, S.A., Huynen, M.A. and Ouzounis, C.A. (2005) The properties of protein family space depend on experimental design. Bioinformatics, 21, 2618–2622.
31. Pushker, R., Mira, A. and Rodriguez-Valera, F. (2004) Comparative genomics of gene-family size in closely related bacteria. Genome Biol., 5, R27.
32. Lynch, M. and Conery, J.S. (2000) The evolutionary fate and consequences of duplicate genes. Science, 290, 1151–1155.
33. Li, W.-H. (1997) Molecular Evolution. Sinauer Associates, Inc., Sunderland Massachusetts, USA.